Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Υλοποίηση Αποθήκης Μεταναστευτικών Δεδομένων – OLAP Ανάλυση – Data mining μοντέλα**
Ονοματεπώνυμο Φοιτητή	**Βασιλική Καρασιώτου**
Πατρώνυμο	**Χρήστος**
Αριθμός Μητρώου	**ΜΠΠΛ/ 06021**
Επιβλέπων	**Γιάννης Θεοδωρίδης, Αναπληρωτής Καθηγητής**

Ημερομηνία Παράδοσης **Απρίλιος 2010**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γιάννης Θεοδωρίδης Αναπλ. Καθηγητής	Δημήτρης Δεσπότης Καθηγητής	Δημήτρης Αποστόλου Λέκτορας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ. Γιάννη Θεοδωρίδη, Αναπλ. Καθηγητή του Τμήματος Πληροφορικής του Πανεπιστήμιου Πειραιώς, για την καθοδήγηση και υποστήριξη κατά την υλοποίηση της παρούσας εργασίας, αλλά και για την πολύ καλή συνεργασία που είχαμε όλο αυτό το διάστημα.

Παράλληλα, θα ήθελα να εκφράσω τις ευχαριστίες μου στον κ. Χρήστο Σκιαδά, Προϊστάμενο της Διεύθυνσης Αναλογιστικών Μελετών και Στατιστικής του ΙΚΑ-ΕΤΑΜ για την προσφορά πρωτογενών δεδομένων. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Γεράσιμο Μαρκέτο, Διδάκτορα του Τμήματος Πληροφορικής του Πανεπιστήμιου Πειραιώς, για τις συμβουλές του.

Τέλος, θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στο συμφοιτητή μου, Πέτρο Σταυγιανουδάκη για τις συζητήσεις μας και τη βοήθειά του κατά την υλοποίηση της παρούσας εργασίας, καθώς και στο σύζυγό μου Παναγιώτη Κρασόπουλο για την προτροπή και υποστήριξή του κατά τη διάρκεια των σπουδών μου στο μεταπτυχιακό πρόγραμμα «Πληροφορική».

Περίληψη

Σκοπός της εργασίας είναι η σχεδίαση και η ανάπτυξη ενός πολυδιάστατου συστήματος συλλογής, αποθήκευσης, επεξεργασίας και αξιοποίησης μεταναστευτικών δεδομένων των ενεργών αλλοδαπών ασφαλισμένων των Ασφαλιστικών Οργανισμών ΙΚΑ-ΕΤΑΜ (Ίδρυμα Κοινωνικών Ασφαλίσεων – Ενιαίο Ταμείο Ασφάλισης Μισθωτών) και ΟΑΕΕ/ΤΕΒΕ (Οργανισμός Ασφάλισης Ελεύθερων Επαγγελματιών / Ταμείο Επαγγελματιών και Βιοτεχνών Ελλάδος). Το σύστημα αξιοποιεί σύγχρονες τεχνολογίες πληροφορικής (Data Warehouses, On-line Analytical Processing/OLAP, Data Mining) για τη διαχείριση των δημογραφικών και οικονομικών δεδομένων που συλλέγουν και αποθηκεύουν οι φορείς κοινωνικής ασφάλισης. Στη συνέχεια, τα δεδομένα αυτά οργανώνονται σε πολυδιάστατα μοντέλα δεδομένων – κύβους όπου εφαρμόζονται κατάλληλες τεχνικές ανάλυσης για ανάκτηση της αποθηκευμένης πληροφορίας και μετατροπή της σε γνώση για λήψη αποφάσεων. Ουσιαστικά, στο περιεχόμενο της εργασίας παρουσιάζονται τα εξής θέματα: α) τα βήματα υλοποίησης της Αποθήκης Μεταναστευτικών Δεδομένων, β) η εφαρμογή λειτουργιών αναλυτικής επεξεργασίας OLAP και MDX ερωτημάτων πάνω σε κύβο, και γ) η εφαρμογή τεχνικών εξόρυξης γνώσης (Classification, Clustering και Association Rules) με τη χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με τη διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases – KDD).

Λέξεις κλειδιά**:** Βάσεις δεδομένων, Αποθήκες Δεδομένων, Λειτουργίες Αναλυτικής Επεξεργασίας Δεδομένων, Τεχνικές Εξόρυξης Γνώσης

.

Abstract

The aim of this master thesis is to design and develop a multidimensional system that collects, stores, processes and utilizes demographic data on actively employed foreigners insured in either of the two largest Social Security Organisations in Greece: IKA-ETAM (Social Insurance Institute – United Insurance Fund for Employees) and OAEE/TEBE (Free-lancers' Employment Organization). The system uses modern technologies of the informatics science such us Data Warehouses, On-line Analytical Processing/OLAP and Data Mining Techniques for the treatment of demographic and economic data that are collected and stored by Social Security Organisations. Data is then organized in multidimensional models; cubes where suitable analysis techniques are applied aiming to retrieve the stored information and to convert it to knowledge for decision-making. This essay summarizes the following issues: a) the steps that where followed in order to implement the Emigrational Data Warehouse, b) the application of OLAP operations and MDX queries in cubes; and c) the application of data mining techniques like Classification, Clustering and Association Rules using algorithms for extraction of information and of patterns that are generated during the process of Knowledge Discovery in Databases – KDD.

Keywords**:** Data Bases, Data Warehouses, On-line Analytical Processing/OLAP Operations, Data Mining Techniques

Περιεχόμενα

1. Εισαγωγή

Το έναυσμα για την εκπόνηση της παρούσας εργασίας προήλθε από τη διακήρυξη ανοιχτού διαγωνισμού του Ινστιτούτου Μεταναστευτικής Πολιτικής (Ι.ΜΕ.ΠΟ) για την υλοποίηση Πληροφοριακού Συστήματος Διαχείρισης και Αξιοποίησης Μεταναστευτικών Δεδομένων, στα πλαίσια εκσυγχρονισμού του, με σύγχρονες τεχνολογίες πληροφορικής και επικοινωνιών για τη διαμόρφωση πολιτικής σε θέματα μετανάστευσης.

Το Ι.ΜΕ.ΠΟ είναι Νομικό Πρόσωπο Ιδιωτικού Δικαίου και τελεί υπό την εποπτεία του Υπουργείο Εσωτερικών, Αποκέντρωσης & Ηλεκτρονικής Διακυβέρνησης. Η αποστολή του Ι.ΜΕ.ΠΟ συνοψίζεται στην έρευνα του μεταναστευτικού φαινομένου και την εκπόνηση μελετών για το σχεδιασμό και την εφαρμογή μιας βιώσιμης και ρεαλιστικής μεταναστευτικής πολιτικής στην Ελλάδα στο πλαίσιο της Ευρωπαϊκής Ένωσης. Παράλληλα, το Ι.ΜΕ.ΠΟ καλείται να αναλάβει το ρόλο του Συμβούλου της Κυβέρνησης για τη διαμόρφωση πολιτικής σε θέματα μετανάστευσης. Στόχος του Ινστιτούτου είναι η σφαιρική αντιμετώπιση του ζητήματος της μετανάστευσης με την οργάνωση ερευνών και την υποβολή προτάσεων που καλύπτουν όλες τις πτυχές της ένταξης των μεταναστών στον Ελλαδικό χώρο. Οι κύριοι φορείς που εμπλέκονται στον κύκλο ζωής ενός αλλοδαπού στην Ελλάδα και που θα αποτελέσουν τις πηγές μεταναστευτικών δεδομένων (πρωτογενών δεδομένων) που θα τροφοδοτήσουν την Αποθήκη Δεδομένων του Πληροφοριακού Συστήματος του Ινστιτούτου είναι οι εξής:

- Υπουργείο Εσωτερικών, Αποκέντρωσης & Ηλεκτρονικής Διακυβέρνησης
- Οι ασφαλιστικοί οργανισμοί
    - Ίδρυμα Κοινωνικών Ασφαλίσεων – Ενιαίο Ταμείο Ασφάλισης Μισθωτών (ΙΚΑ-ΕΤΑΜ)
    - Οργανισμός Γεωργικών Ασφαλίσεων (ΟΓΑ)
    - Οργανισμός Ασφάλισης Ελεύθερων Επαγγελματιών / Ταμείο Επαγγελματιών και Βιοτεχνών Ελλάδος (ΟΑΕΕ/ΤΕΒΕ)

Το ΙΚΑ_ΕΤΑΜ είναι ο μεγαλύτερος ασφαλιστικός οργανισμός της χώρας. Μέσω της ανάπτυξης Ολοκληρωμένου Πληροφοριακού Συστήματος (ΟΠΣ-ΙΚΑ-ΕΤΑΜ) που καλύπτει το σύνολο των λειτουργιών και διαδικασιών του Ιδρύματος, υλοποιεί ένα πρόγραμμα εκσυγχρονισμού της οργάνωσης και των διαδικασιών του. Το ΟΠΣ-ΙΚΑ λειτουργεί από το 2002 και στο Μητρώο του υπάρχουν 550.000 με 600.000 απογεγραμμένοι αλλοδαποί (υπήκοοι χωρών-μελών της Ευρωπαϊκής Ένωσης και τρίτων χωρών) με δικαίωμα περίθαλψης (δηλαδή άνω των 150 ημερομισθίων ανά έτος). Επίσης, ο ΟΑΕΕ προήλθε από την ενοποίηση των Ασφαλιστικών Ταμείων ΤΕΒΕ (Ταμείο Επαγγελματιών και Βιοτεχνών Ελλάδος), ΤΑΕ (Ταμείο Ασφάλισης Εμπόρων) και ΤΣΑ (Ταμείο Συντάξεων Αυτοκινητιστών) με το νόμο 2676/1999 και περιλαμβάνει τους κλάδους Σύνταξης και Υγείας. Το 2004 υπήρχαν στον ΟΑΕΕ περίπου 855.000 ασφαλισμένοι εκ των οποίων 570.000 ανήκαν στο ΤΕΒΕ, 205.000 στο ΤΑΕ και 80.000 στο ΤΣΑ. Επομένως, η αποτελεσματική αποθήκευση και επεξεργασία του τεράστιου όγκου πληροφοριών που συγκεντρώνουν οι παραπάνω φορείς, με την υλοποίηση Πληροφοριακού Συστήματος, έχει ως σκοπό την άμεση αξιοποίηση πρωτογενών δεδομένων για τους μετανάστες κατά τρόπο οργανωμένο και συστηματικό ώστε να υποστηριχθούν οι έρευνες και οι μελέτες του Ινστιτούτου που αφορούν μεταξύ άλλων την απασχόληση, την κοινωνική ενσωμάτωση, την πρόσβαση σε υγειονομικές και κοινωνικές υπηρεσίες, την εκπαίδευση, τη στέγη και τα οικιστικά θέματα, την ενεργό ανάμιξη και συμμετοχή των μεταναστών στην κοινωνία κ.α.

Ωστόσο, ο λογικός και φυσικός Πληροφοριακού Συστήματος Διαχείρισης και Αξιοποίησης Μεταναστευτικών Δεδομένων που προέρχονται από ετερογενείς πηγές πληροφόρησης (Υπουργείο Εσωτερικών, Αποκέντρωσης και Ηλεκτρονικής Διακυβέρνησης, ΙΚΑ-ΕΤΑΜ, ΟΓΑ, ΟΑΕΕ/ΤΕΒΕ) αποτελεί ένα μεγαλεπήβολο και εξαιρετικά δύσκολο έργο που μπορεί να αντιμετωπισθεί αποτελεσματικά μόνο με τη χρήση προηγμένων συστημάτων διαχείρισης βάσεων δεδομένων και υποστήριξης λήψης αποφάσεων.

Στα πλαίσια υλοποίησης της παρούσας εργασίας παρουσιάζεται η αξιοποίηση νέων τεχνολογιών διαχείρισης και ανάλυσης του τεράστιου όγκου ψηφιακής πληροφορίας που αποθηκεύεται σε βάσεις δεδομένων. Ουσιαστικά, υλοποιήθηκε σε μικρή κλίμακα ένα πολυδιάστατο σύστημα συλλογής, αποθήκευσης, επεξεργασίας και αξιοποίησης

μεταναστευτικών δεδομένων των ενεργών αλλοδαπών ασφαλισμένων των Ασφαλιστικών Οργανισμών ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ/ΤΕΒΕ με την εφαρμογή σύγχρονων τεχνολογιών πληροφορικής, όπως είναι οι αποθήκες δεδομένων (data warehouses), οι λειτουργίες αναλυτικής επεξεργασίας δεδομένων (OLAP) και οι τεχνικές εξόρυξης γνώσης (data mining techniques).

Η δομή της μεταπτυχιακής διατριβής αποτελείται από τα εξής κεφάλαια: Στο κεφάλαιο 2 γίνεται αναφορά στις σύγχρονες τεχνολογίες πληροφορικής που έχουν αναπτυχθεί και εφαρμόζονται για τη διαχείριση του τεράστιου όγκου των αποθηκευμένων δεδομένων καθώς και για την αναλυτική επεξεργασία τους με σκοπό την εξαγωγή χρήσιμης και άμεσα αξιοποιήσιμης γνώσης πληροφορίας από τα δεδομένα. Ουσιαστικά, στόχος του κεφαλαίου αυτού είναι η εξοικείωση του αναγνώστη με τις σχετικές έννοιες και ορολογίες ώστε να κατανοήσει καλύτερα τις μεθοδολογίες και τις αντίστοιχες διαδικασίες που εφαρμόστηκαν στα επόμενα κεφάλαια της εργασίας. Στο κεφάλαιο 3 παρουσιάζονται τα βήματα που ακολουθήθηκαν για την υλοποίηση της Αποθήκης Μεταναστευτικών Δεδομένων, τα οποία περιλαμβάνουν τον ορισμό της αρχιτεκτονικής του συστήματος, την υλοποίηση βάσης δεδομένων, την παραγωγή και προπαρασκευή των δεδομένων και τέλος την επιλογή του μοντέλου – σχήματος της Αποθήκης. Για τους σκοπούς της υλοποίησης της Βάσης και της Αποθήκης Δεδομένων χρησιμοποιήθηκε το Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) του Microsoft SQL Server 2005. Στα κεφάλαια 4, 5 παρουσιάζονται τα πολυδιάστατα μοντέλα δεδομένων – κύβοι που δημιουργήθηκαν πάνω στα δεδομένα της Αποθήκης, ώστε να εφαρμοστούν σε αυτά λειτουργίες αναλυτικής επεξεργασίας δεδομένων και τεχνικές εξόρυξης γνώσης. Το κεφάλαιο 4 καλύπτει την προπαρασκευή της Αποθήκης Μεταναστευτικών Δεδομένων και το σχεδιασμό του κύβου ενώ παράλληλα δίνονται παραδείγματα εκτέλεσης λειτουργιών OLAP και καταγράφονται τα σχετικά συμπεράσματα. Το κεφάλαιο 5 καλύπτει την προπαρασκευή της Αποθήκης Μεταναστευτικών Δεδομένων και το σχεδιασμό των κύβων πάνω στους οποίους υλοποιήθηκαν μοντέλα εξόρυξης γνώσης με την εφαρμογή τεχνικών Κατηγοριοποίησης (Classification), Συσταδοποίησης (Clustering) και Εύρεσης Κανόνων Συσχετίσεων (Association Rules). Για κάθε τεχνική υλοποιήθηκαν δύο μοντέλα σε δύο διαφορετικούς κύβους και σχολιάστηκαν σε αντιπαραβολή τα εξαγόμενα σχετικά συμπεράσματα. Το εργαλείο που χρησιμοποιήθηκε για τους σκοπούς της OLAP ανάλυσης και των τεχνικών εξόρυξης γνώσης είναι το Analysis Services του Microsoft SQL Server 2005. Στο κεφάλαιο 6 παρουσιάζονται τα τελικά συμπεράσματα από την εκπόνηση της παρούσας εργασίας ενώ παράλληλα δίνονται κάποιες παρατηρήσεις σχετικά με την ευελιξία των πολυδιάστατων μοντέλων δεδομένων κατά την εφαρμογή πάνω σε αυτά τεχνικών εξόρυξης γνώσης με το εργαλείο Analysis Services του Microsoft SQL Server 2005. Τέλος, στο κεφάλαιο 7 παρουσιάζεται η σχετική βιβλιογραφία που χρησιμοποιήθηκε ενώ ακολουθεί και το κεφάλαιο 8 το οποίο αποτελείται από τρία παραρτήματα, που συμπληρώνουν την περιγραφή των σταδίων υλοποίησης της μεταπτυχιακής διατριβής. Πιο αναλυτικά, στο παράρτημα Α δίνονται τα SQL και MDX Scripts που χρησιμοποιήθηκαν στις βάσεις και στους κύβους, στο παράρτημα Β παρατίθενται επίσημα δημοσιευμένα έντυπα και στατιστικά που χρησιμοποιήθηκαν ως πρωτογενείς πηγές άντλησης πληροφορίας, ενώ στο τελευταίο παράρτημα Γ δίνεται ο πηγαίος κώδικας των τριών προγραμμάτων που υλοποιήθηκαν σε Matlab.

2. Αποθήκες Δεδομένων και Εξόρυξη Γνώσης

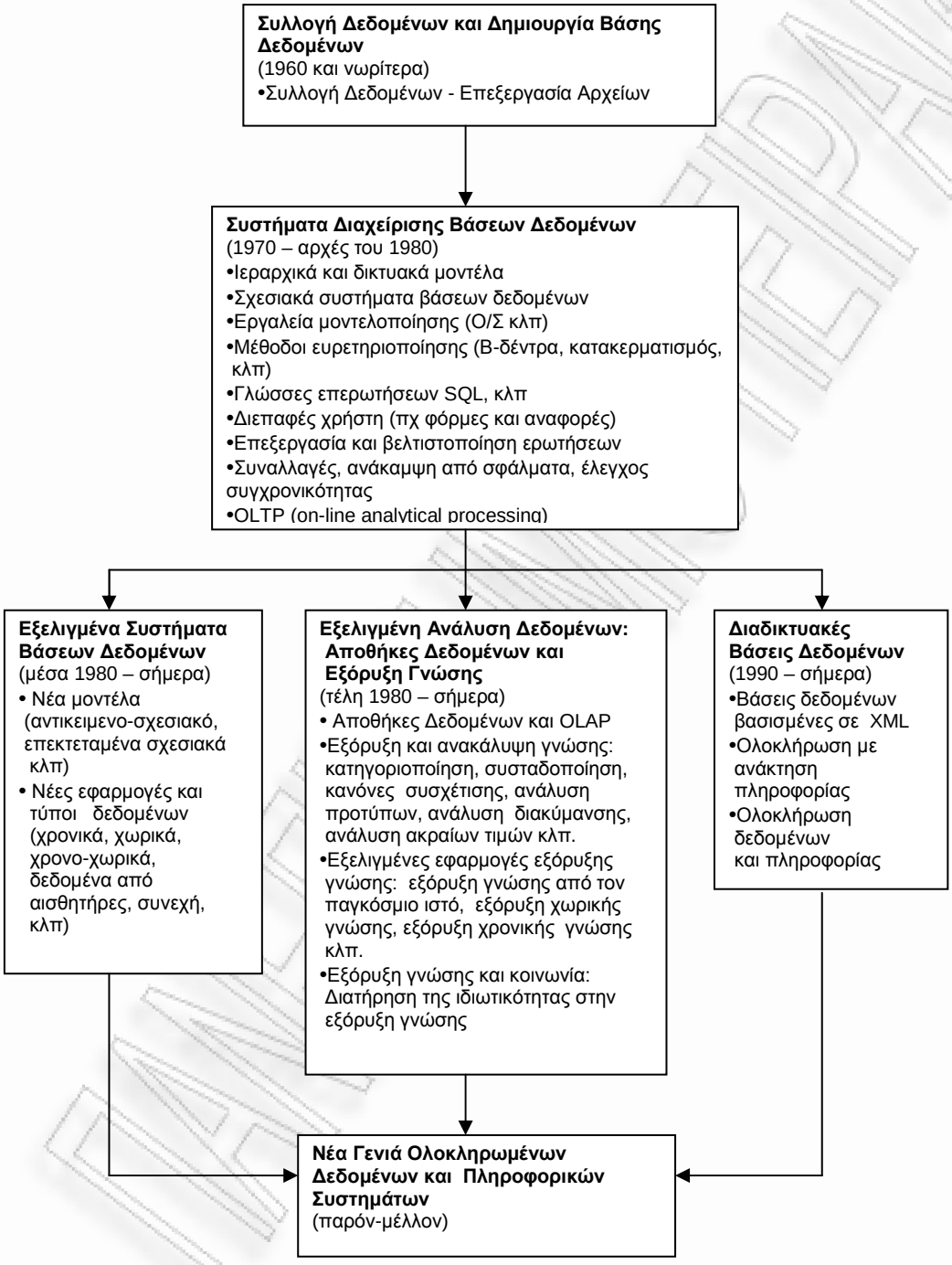
Η σύγχρονη εποχή συχνά αναφέρεται ως εποχή της πληροφορίας λόγω του μεγάλου όγκου των δεδομένων που παράγονται και κατ'επέκταση αποθηκεύονται χάρη στη γρήγορη και φθηνή τεχνολογία αποθήκευσης. Η έλευση των υπολογιστών και των μέσων μαζικής αποθήκευσης ψηφιακής πληροφορίας κατέστησε δυνατή τη συλλογή και αποθήκευση κάθε είδους δεδομένων όχι όμως και εύκολη τη διαχείριση των δεδομένων αυτών. Το πρόβλημα της διαχείρισης της πληθώρας δεδομένων που προέρχονται από ετερογενείς πηγές αντιμετωπίστηκε με τη δημιουργία δομημένων βάσεων δεδομένων καθώς και συστημάτων διαχείρισης βάσεων δεδομένων (DBMS). Η παραγωγή ισχυρών συστημάτων διαχείρισης βάσεων δεδομένων βοήθησε σημαντικά στην ανάπτυξη πληροφοριακών συστημάτων που καλύπτουν τις λειτουργικές ανάγκες οργανισμών και επιχειρήσεων. Ο πυρήνας κάθε πληροφοριακού συστήματος είναι η βάση δεδομένων του και η σωστή σχεδίαση, ανάπτυξη και λειτουργία της βάσης εξασφαλίζει την επιτυχία του πληροφοριακού συστήματος. Τα πληροφορικά συστήματα διακρίνονται σε συστήματα επεξεργασίας μεγάλου αριθμού δοσοληψιών των δεδομένων ενός οργανισμού (On-line transaction processing – OLTP) και σε συστήματα στήριξης αποφάσεων (Decision Support Systems – DSS) που βοηθούν τα στελέχη των οργανισμών στη λήψη αποφάσεων.

Μια από τις βασικές απαιτήσεις των συστημάτων στήριξης αποφάσεων είναι η αποδοτική πρόσβαση στα δεδομένα. Ωστόσο, αυτό δεν είναι πάντοτε εφικτό καθώς οι βάσεις δεδομένων έχουν μεγάλο υπολογιστικό φορτίο και έχουν σχεδιαστεί για την εκτέλεση συγκεκριμένων λειτουργιών. Επίσης, πολλές φορές οι βάσεις έχουν υλοποιηθεί με τη χρήση τεχνολογίας που θεωρείται πλέον παρωχημένη (π.χ. αρχεία COBOL) και επομένως εφαρμογές που χρησιμοποιούν μοντέρνα τεχνολογία δεν μπορούν να διαχειριστούν πληροφορία που προέρχεται από βάση δεδομένων παλαιάς τεχνολογίας. Τέλος, κάθε σύστημα στήριξης αποφάσεων εκτελεί μεγάλο αριθμό ερωτήσεων με αποτέλεσμα να δεσμεύει αρκετούς από τους πόρους του συστήματος διαχείρισης της βάσης δεδομένων. Αυτό έχει ως συνέπεια τη μείωση της απόδοσης του συστήματος το οποίο αρχικά σχεδιάστηκε για να λειτουργεί συνεχώς και η βάση του να εξυπηρετεί μεγάλο όγκο δοσοληψιών. Επομένως, γίνεται σαφές ότι είναι εξαιρετικά δύσκολη ή και πρακτικά αδύνατη η χρήση βάσεων δεδομένων πληροφορικών συστημάτων από συστήματα στήριξης αποφάσεων. Η λύση στο πρόβλημα δόθηκε με την ανάπτυξη των "Αποθηκών Δεδομένων".

"Οι Αποθήκες Δεδομένων (Data Warehouses) αποτελούν θεματο-κεντρικά (subject - oriented), συγκεντρωμένα (integrated), με χρονική διάσταση (time - variable), μη ευμετάβλητα (non – volatile) συστήματα διαχείρισης πληροφοριακών δεδομένων για την υποστήριξη των διαδικασιών λήψης αποφάσεων" [10]. Μια Αποθήκη Δεδομένων αντλεί δεδομένα από βάσεις δεδομένων πληροφοριακών συστημάτων αλλά και από άλλες πηγές δεδομένων όπως αρχεία και δεδομένα που προέρχονται από εξωτερικές πηγές. Στη συνέχεια τα δεδομένα αυτά οργανώνονται στην Αποθήκη με κατάλληλες δομές που ανταποκρίνονται στις απαιτήσεις των αναλυτών – χρηστών των συστημάτων στήριξης αποφάσεων και παρέχουν αποδοτική πρόσβαση στα δεδομένα, χωρίς την παρουσία των προαναφερθέντων προβλημάτων. Οι Αποθήκες Δεδομένων παρέχουν τη δυνατότητα για συνεχή Αναλυτική Επεξεργασία (On-line Analytical Processing – OLAP) συγκεντρωμένης ιστορικής πληροφορίας χρήσιμης για την υποστήριξη αποφάσεων και τις εφαρμογές στρατηγικού σχεδιασμού.

Ωστόσο, η εύκολη ανάκτηση της πληροφορίας που εξασφαλίζουν οι Αποθήκες Δεδομένων δεν είναι αρκετή για τη λήψη αποφάσεων. Οι τεράστιες συλλογές ποικίλλων δεδομένων, όπως απλές αριθμητικές μετρήσεις, κείμενα αλλά και πιο σύνθετη πληροφορία όπως χωρικά δεδομένα, χρονικά δεδομένα και δεδομένα που αντλούνται από τον Παγκόσμιο Ιστό, δημιουργούν νέες ανάγκες για καλύτερες επιλογές διαχείρισης των δεδομένων. Τέτοιες ανάγκες είναι: η αυτόματη σύνοψη των δεδομένων, η εξαγωγή της ουσίας της αποθηκευμένης πληροφορίας καθώς και η ανακάλυψη προτύπων από ακατέργαστα δεδομένα. Οι όροι ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases – KDD) και εξόρυξη γνώσης από δεδομένα (Data Mining – DM) συχνά αναφέρονται στην ίδια έννοια, δηλαδή στη διαδικασία ανακάλυψης χρήσιμων, συνήθως κρυμμένων προτύπων από τα δεδομένα.

Σε σχέση με την ιστορία των βάσεων δεδομένων όπως φαίνεται στο παρακάτω σχήμα, η διαδικασία ανακάλυψης γνώσης που περιλαμβάνει το σχεδιασμό Αποθηκών Δεδομένων, τη συλλογή και την προεπεξεργασία των δεδομένων, την εξόρυξη γνώσης, την επιλογή μοντέλου ή συνδυασμού μοντέλων, την αξιολόγηση και τελικά την ενοποίηση και χρησιμοποίηση της εξαγόμενης γνώσης για τη λήψη αποφάσεων, είναι πολύ καινούργια.



Σχήμα 2-1. Η εξέλιξη της τεχνολογίας διαχείρισης συστημάτων βάσεων δεδομένων.

2.1 Αποθήκες Δεδομένων

Οι Αποθήκες Δεδομένων γενικεύουν και ενοποιούν τα δεδομένα σε πολυδιάστατο χώρο. Ουσιαστικά αποτελούν ένα σύνολο τεχνολογιών που παρέχει τη δυνατότητα στους αναλυτές ενός οργανισμού – επιχείρησης να σχεδιάσουν την πολιτική του έχοντας αποδοτική πρόσβαση στα δεδομένα του οργανισμού – επιχείρησης. Η υλοποίηση μιας Αποθήκης Δεδομένων περιλαμβάνει το σχεδιασμό μιας κεντρικής βάσης δεδομένων με σκοπό τη συγκέντρωση ετερογενών πηγών πληροφοριών σε μία τοποθεσία και παράλληλα την αποφυγή σύγκρουσης μεταξύ συστημάτων επεξεργασίας συναλλαγών (OLTP) και συστημάτων αναλυτικής επεξεργασίας δεδομένων (OLAP). Η σχεδίαση Αποθηκών Δεδομένων έχει σαν στόχο την αποδοτική απάντηση πολύπλοκων ερωτήσεων που δημιουργούνται κατά την αναλυτική επεξεργασία των δεδομένων και συντελεί στην αύξηση της αποδοτικότητας των εφαρμογών για τη λήψη αποφάσεων και τη χάραξη στρατηγικού σχεδιασμού. Η δημιουργία και η συντήρηση μιας Αποθήκης Δεδομένων είναι μια πολύπλοκη διαδικασία και εξαρτάται από τους στόχους που θέτει κάθε οργανισμός κατά την αναλυτική επεξεργασία των δεδομένων του. Πολλοί οργανισμοί επιδιώκουν τη δημιουργία Αποθήκης Δεδομένων στην οποία να συγκεντρώνεται η αναλυτική πληροφορία από τις δραστηριότητες του οργανισμού γεγονός που αυξάνει σημαντικά το κόστος υλοποίησης της Αποθήκης. Ενίοτε, η Αποθήκη Δεδομένων ενός οργανισμού συμπληρώνεται από εξειδικευμένα θεματικά υποσύνολα – επιμέρους συλλογές δεδομένων (data marts) για περαιτέρω απόδοση των OLAP εφαρμογών, καθώς πρόκειται για πιο ευέλικτα συστήματα στη δημιουργία τους που όμως δεν παρέχουν ενιαία λύση, ενώ η μακροχρόνια χρήση τους δημιουργεί προβλήματα.

Στη συνέχεια, θα μελετήσουμε τι ακριβώς είναι μια Αποθήκη Δεδομένων, την αρχιτεκτονική και το σχήμα της και τις λειτουργικές της διαδικασίες, καθώς όλο και περισσότεροι οργανισμοί υλοποιούν Αποθήκες Δεδομένων για την ανάλυση των δεδομένων τους. Ιδιαίτερα, θα μελετήσουμε τον κύβο δεδομένων (data cube), δηλαδή το πολυδιάστατο μοντέλο δεδομένων για τις Αποθήκες Δεδομένων και την αναλυτική επεξεργασία των δεδομένων καθώς επίσης και τις βασικές λειτουργίες OLAP (τεμαχισμός – slice, κομμάτιασμα – dice, συσσώρευση – roll-up, εμβάθυνση – drill-down). Η αναλυτική αναφορά στην τεχνολογία που έχει αναπτυχθεί γύρω από τις Αποθήκες Δεδομένων θα συνεχιστεί στις επόμενες παραγράφους με την αναλυτική παρουσίαση των τεχνικών εξόρυξης γνώσης ώστε να ολοκληρωθεί το θεωρητικό υπόβαθρο πάνω στο οποίο στηρίχθηκε η σχεδίαση και υλοποίηση της παρούσας εργασίας.

2.1.1 Η Αποθήκη Δεδομένων – Τι είναι και ποιες οι διαφορές της από τις λειτουργικές βάσεις δεδομένων

Η χρησιμοποίηση της τεχνολογίας των Αποθηκών Δεδομένων παρέχει στους αναλυτές και τα διευθυντικά στελέχη των επιχειρήσεων τεχνικές και εργαλεία για τη συστηματική οργάνωση, κατανόηση και τελικά χρήση των δεδομένων για τη χάραξη στρατηγικού σχεδιασμού. Τα συστήματα Αποθηκών Δεδομένων αποτελούν χρήσιμα εργαλεία στο σύγχρονο ανταγωνιστικό και γρήγορα εξελισσόμενο κόσμο. Τα τελευταία χρόνια η ανάπτυξη και λειτουργία Αποθηκών Δεδομένων είναι πολύ σημαντική για τη λειτουργία οργανισμών και επιχειρήσεων καθώς οι περισσότεροι πιστεύουν ότι στη σύγχρονη ανταγωνιστική βιομηχανία τέτοια συστήματα παρέχουν σημαντικά οφέλη, με αποτέλεσμα να επενδύονται τεράστια ποσά σε αντίστοιχες δραστηριότητες. Αυτός είναι και ο λόγος που όλες οι μεγάλες εταιρείες του χώρου των Βάσεων Δεδομένων και των πληροφοριακών συστημάτων αναπτύσσουν και προτείνουν προϊόντα στο χώρο των Αποθηκών Δεδομένων και στα επόμενα χρόνια αναμένονται ακόμα μεγαλύτερες επενδύσεις σε αντίστοιχη τεχνολογία.

Οι Αποθήκες Δεδομένων έχουν οριστεί με ποικίλους τρόπους γεγονός που καθιστά δύσκολη τη διατύπωση ενός αυστηρού ορισμού. Γενικά, με τον όρο Αποθήκες Δεδομένων (Data Warehouses) χαρακτηρίζεται ένα σύνολο τεχνολογιών που υποστηρίζουν την αποδοτική πρόσβαση στα δεδομένα ενός οργανισμού και την επεξεργασία τους με σκοπό τη σχεδίαση της πολιτικής του. Ουσιαστικά, μια Αποθήκη Δεδομένων είναι μια βάση δεδομένων που υποστηρίζει αποφάσεις και συντηρείται ξεχωριστά από τη λειτουργική βάση δεδομένων (Operational Database) ενός οργανισμού.

Ένας γενικός ορισμός δίδεται ως εξής: “Οι Αποθήκες Δεδομένων (Data Warehouses) αποτελούν θεματο-κεντρικά (subject - oriented), συγκεντρωμένα (integrated), με χρονική διάσταση (time - variant), μη ευμετάβλητα (non – volatile) συστήματα διαχείρισης πληροφοριακών δεδομένων για την υποστήριξη των διαδικασιών λήψης αποφάσεων” [10]. Ο σύντομος αλλά ωστόσο πολύ περιεκτικός αυτός ορισμός συνοψίζει τα κύρια χαρακτηριστικά μιας Αποθήκης Δεδομένων. Οι τέσσερις λέξεις κλειδιά, θεματο-κεντρικά, συγκεντρωμένα, με χρονική διάσταση και μη ευμετάβλητα, διαφοροποιούν τις Αποθήκες Δεδομένων από άλλα συστήματα αποθήκευσης δεδομένων, όπως για παράδειγμα τα σχεσιακά συστήματα βάσεων δεδομένων, τα συστήματα επεξεργασίας συναλλαγών και τα συστήματα αρχείων. Ειδικότερα:

- Θεματο-κεντρικά: Μια Αποθήκη Δεδομένων οργανώνεται γύρω από συγκεκριμένα θέματα, όπως πελάτες, προμηθευτές, προϊόντα, πωλήσεις και επικεντρώνεται στη μοντελοποίηση και στην ανάλυση των δεδομένων για τη λήψη αποφάσεων. Δεδομένα ή πλευρές θεμάτων που χρησιμεύουν στη διαδικασία υποστήριξης αποφάσεων δεν συμπεριλαμβάνονται στην Αποθήκη.
- Συγκεντρωμένα: Μια Αποθήκη Δεδομένων συνήθως κατασκευάζεται με συγκέντρωση πολλαπλών ετερογενών πηγών, όπως σχεσιακές βάσεις δεδομένων, αρχεία κλπ. Στη συνέχεια εφαρμόζονται τεχνικές καθαρισμού και ολοκλήρωσης δεδομένων (π.χ. δομές κωδικοποίησης, μέτρα των χαρακτηριστικών, συμβάσεις ονοματολογίας κλπ.) για την εξασφάλιση συνέπειας των ετερογενών δεδομένων.
- Με χρονική διάσταση: Τα δεδομένα αποθηκεύονται για να παρέχουν ιστορική πληροφορία (π.χ. για τα τελευταία 5-10 χρόνια). Γενικά, η έννοια του χρόνου είναι αναπόσπαστο τμήμα μιας Αποθήκης Δεδομένων.
- Μη ευμετάβλητα: Μια Αποθήκη Δεδομένων αποθηκεύεται ξεχωριστά από τη λειτουργική βάση δεδομένων και τα δεδομένα της δεν υπόκεινται σε τροποποιήσεις (π.χ. επεξεργασία συναλλαγών, ανάνηψη, έλεγχος συνδρομικότητας – concurrency control). Στις Αποθήκες Δεδομένων υπάρχει μόνο η λειτουργία της φόρτωσης είτε πλήρως (full loading) είτε αυξητικά (incremental loading).

Η κατασκευή και η συντήρηση μιας Αποθήκης Δεδομένων είναι μια πολύπλοκη διαδικασία και εξαρτάται από τις ανάγκες κάθε οργανισμού ή επιχείρησης. Πολλοί οργανισμοί επιδιώκουν να δημιουργήσουν μια Αποθήκη Δεδομένων που θα περιέχει αναλυτικά δεδομένα από όλες τις δραστηριότητες του οργανισμού. Ένα τέτοιο εγχείρημα απαιτεί πολύ μεγάλο κόστος για να επιτύχει. Συχνά, μια Αποθήκη Δεδομένων συμπληρώνεται από εξειδικευμένα θεματικά υποσύνολα ή αλλιώς από Επιμέρους Συλλογές Δεδομένων (data marts) για επιπλέον απόδοση των OLAP εφαρμογών. Οι Συλλογές Δεδομένων είναι πιο ευέλικτα συστήματα στη δημιουργία τους, τα οποία έχουν ως στόχο την ολοκλήρωση (integration) ετερογενών πηγών πληροφοριών με τη συγκέντρωση όλης της ενδιαφέρουσας πληροφορίας σε μια τοποθεσία, και την αποφυγή της σύγκρουσης μεταξύ OLTP (on-line transaction processing) και OLAP (on-line analytical processing) συστημάτων ώστε να εξασφαλίζεται η απόδοση των εφαρμογών και η διαθεσιμότητα του συστήματος. Ωστόσο, η μακρόχρονη χρήση τους δημιουργεί προβλήματα.

Επειδή οι περισσότεροι άνθρωποι είναι εξοικειωμένοι με τα εμπορικά συστήματα σχεσιακών βάσεων δεδομένων, είναι αρκετά εύκολο να κατανοήσουμε τι είναι μια Αποθήκη Δεδομένων συγκρίνοντας τα συστήματα OLTP/OLAP.

Ένα Σύστημα Επεξεργασίας Συναλλαγών (OLTP) αποτελεί ένα πλήρες σύστημα που περιέχει εργαλεία για τον προγραμματισμό των εφαρμογών, την εκτέλεση και τη διαχείριση των συναλλαγών. Είναι μια εφαρμογή που δουλεύει συνεχώς, εξελίσσεται συνεχώς, είναι συνήθως κατανεμημένη και περιλαμβάνει μια βάση δεδομένων, κάποιο δίκτυο και τα αντίστοιχα προγράμματα για την εφαρμογή. Από την άλλη πλευρά ένα Σύστημα Αναλυτικής Επεξεργασίας Συναλλαγών (OLAP) παρέχει ευέλικτη, υψηλής απόδοσης πρόσβαση και ανάλυση μεγάλου όγκο σύνθετων δεδομένων από διαφορετικές εφαρμογές, συμμετοχή αθροιστικών και ιστορικών δεδομένων σε πολύπλοκες ερωτήσεις, μεταβολή της “οπτικής γωνίας” παρουσίασης των δεδομένων (π.χ., από πωλήσεις ανά περιοχή -> πωλήσεις ανά τμήμα κλπ.), συμμετοχή πολύπλοκων υπολογισμών (π.χ. στατιστικές συναρτήσεις) και γρήγορες απαντήσεις σε οποιαδήποτε χρονική στιγμή τεθεί ένα ερώτημα (On-Line). Στη συνέχεια παρουσιάζονται αναλυτικά τα βασικά χαρακτηριστικά που διαφοροποιούν τα OLTP συστήματα από τα OLAP ενώ στον Πίνακα 2-1 που ακολουθεί γίνεται μια σύνοψη των διαφορών.

- Χρήστες και προσανατολισμός του συστήματος: Ένα OLTP σύστημα προσανατολίζεται στις απαιτήσεις του πελάτη και χρησιμοποιείται από διοικητικούς υπαλλήλους και

διαχειριστές της βάσης δεδομένων του οργανισμού. Ένα OLAP σύστημα προσανατολίζεται στις απαιτήσεις της αγοράς και χρησιμοποιείται από διευθυντικά στελέχη και αναλυτές.

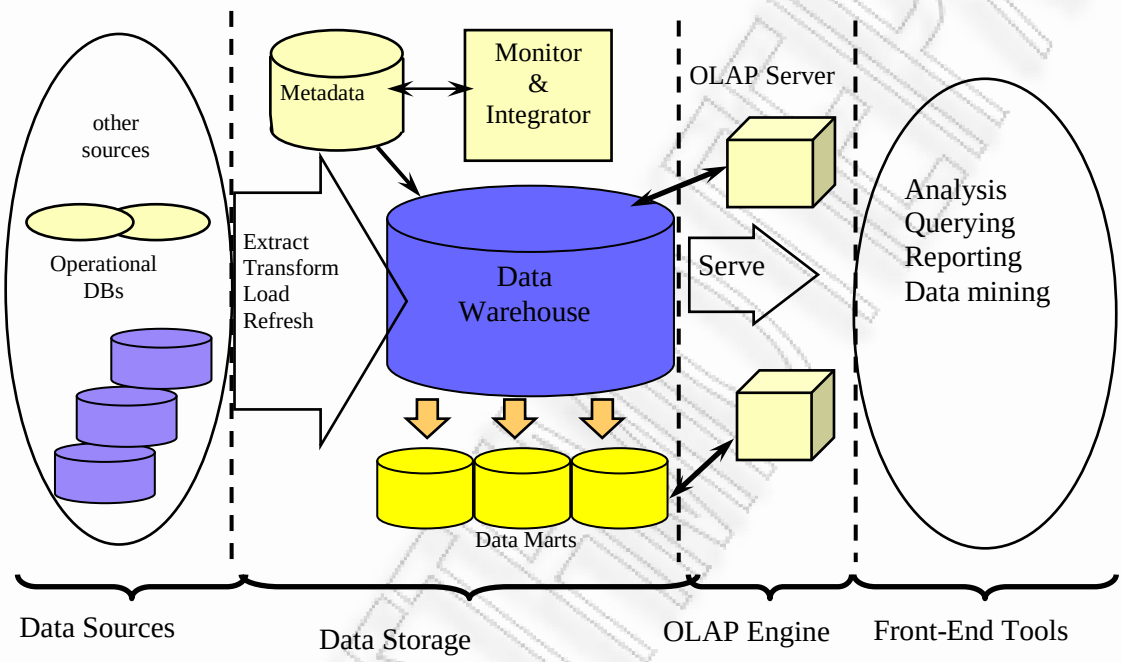
- Περιεχόμενα Δεδομένων: Ένα OLTP σύστημα διαχειρίζεται τρέχοντα – καθημερινά δεδομένα μεγάλης λεπτομέρειας τα οποία εύκολα μπορούν να αναζητηθούν και απαντούν σε απλές ερωτήσεις. Ένα OLAP σύστημα διαχειρίζεται μεγάλες ποσότητες ιστορικής πληροφορίας και παρέχει αποδοτική πρόσβαση στα δεδομένα για λήψη αποφάσεων.
- Σχεδιασμός βάσης δεδομένων: Ένα OLTP σύστημα σχεδιάζεται για να διατηρεί την ακεραιότητα των δεδομένων και να εξασφαλίζει ταχύτητα στην αποθήκευση των καθημερινών συναλλαγών του οργανισμού, επομένως η βάση δεδομένων του συστήματος είναι κανονικοποιημένη βάσει κάποιου μοντέλου Οντοτήτων – Συσχετίσεων (Entity - Relationship model). Ένα OLAP σύστημα σχεδιάζεται για να παρέχει ταχύτητα στην ανάλυση και η βάση δεδομένων του συστήματος είναι από-κανονικοποιημένη βάσει κάποιου μοντέλου αστέρα ή χιονονιφάδας (star/snowflake schema αντίστοιχα θα οποία θα αναλυθούν εκτενώς στην παράγραφο 2.1.3) καθώς οι εφαρμογές OLAP επιταχύνονται αν τα δεδομένα οργανωθούν με μη παραδοσιακούς τρόπους.
- Πρότυπα πρόσβασης (access patterns): Τα δεδομένα ενός OLTP συστήματος υπόκεινται σε λειτουργίες τροποποίησης (π.χ. επεξεργασία συναλλαγών, ανάνηψη, έλεγχος συνδρομικότητας). Από την άλλη πλευρά, τα OLAP συστήματα περιέχουν ιστορική πληροφορία που δεν μεταβάλλεται και επομένως η πρόσβαση σε αυτά επιτρέπει λειτουργίες μόνο για ανάγνωση (read – only).

Πίνακας 2-1. Σύγκριση συστημάτων OLTP/OLAP

	OLTP	OLAP
Δομή	Files/DBMS's	RDBMS
Πρόσβαση	SQL/COBOL/…	SQL & επεκτάσεις
Ανάγκες που καλύπτουν	Αυτοματισμός καθημερινών εργασιών	Άντληση και επεξεργασία πληροφορίας για χάραξη στρατηγικής
Τύπος Δεδομένων	Λεπτομερή, Λειτουργικά	Συνοπτικά, αθροιστικά
Όγκος Δεδομένων	από 100MB έως GB	από 100GB έως TB
Φύση Δεδομένων	Δυναμικά, τρέχοντα	Στατικά, ιστορικά
I/O Τύποι	Περιορισμένο I/O συχνές αναζητήσεις στο δίσκο	Εκτεταμένο I/O συχνές σαρώσεις του δίσκου
Τροποιήσεις	Συνεχείς	Περιοδικές ενημερώσεις
Μέτρηση Απόδοσης	Μέσος Ρυθμός Αποθήκευσης Εγγραφών - Troughput	Χρόνος Απόκρισης
Φόρτος	Συναλλαγές με πρόσβαση λίγων εγγραφών	Ερωτήσεις που σαρώνουν εκατομμύρια εγγραφών
Σχεδίαση Βάσης Δεδομένων	Κατευθυνόμενη από εφαρμογή	Κατευθυνόμενη από περιεχόμενο
Τυπικοί Χρήστες	Χαμηλόβαθμοι Υπάλληλοι, π.χ. διοικητικοί υπάλληλοι, διαχειριστές βάσης δεδομένων	Υψηλόβαθμοι Υπάλληλοι, π.χ. διευθυντικά στελέχη, αναλυτές
Χρήση	Μέσω προκατασκευασμένων φορμών	Ad-hoc
Αριθμός Χρηστών	Χιλιάδες	Δεκάδες
Εστίαση	Εισαγωγή Δεδομένων	Εξαγωγή Πληροφοριών

2.1.2 Η Αρχιτεκτονική της Αποθήκης Δεδομένων

Η επιλογή της αρχιτεκτονικής για μια Αποθήκη Δεδομένων πρέπει να ικανοποιεί τις συγκεκριμένες ανάγκες του οργανισμού για τις οποίες δημιουργήθηκε ώστε να εξασφαλίζεται η διαθεσιμότητα και η αποδοτικότητα του συστήματος. Γενικά, η αρχιτεκτονική μιας Αποθήκης Δεδομένων είναι όπως παρουσιάζεται στο Σχήμα 2-2 όπου σημειώνονται τα βασικά δομικά στοιχεία της Αποθήκης, η διασύνδεση των στοιχείων τους και η ροή των δεδομένων.



Σχήμα 2-2. Αρχιτεκτονική Αποθήκης Δεδομένων

Τα κύρια δομικά μέρη της αρχιτεκτονικής μιας Αποθήκης Δεδομένων είναι τα παρακάτω:

Πηγές δεδομένων (Data sources): Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα. Τα συστήματα διαχείρισης Αποθηκών Δεδομένων αντλούν δεδομένα από διάφορες ετερογενείς πηγές, όπως για παράδειγμα:

- Βάσεις δεδομένων των συστημάτων του οργανισμού.
- Εξωτερικές πηγές πληροφοριών, δηλαδή, πληροφορίες που παρέχονται από πληροφοριακά συστήματα και στα οποία ο οργανισμός έχει πρόσβαση.
- Αρχεία εφαρμογών και αρχεία κειμένου.

ETL (Extract-Transform-Load) εφαρμογές: Εφαρμογές που εκτελούν τις διαδικασίες εξαγωγής, μεταφοράς, μετασχηματισμού, καθαρισμού και φόρτωσης των δεδομένων από τις πηγές στην Αποθήκη Δεδομένων. Πιο αναλυτικά οι παραπάνω εφαρμογές αυτοματοποιούν διαδικασίες όπως:

- Εξαγωγή δεδομένων από τις πηγές.
- Καθαρισμό των δεδομένων με την διάγνωση πιθανών ασυνεπειών και τη μεταφορά μόνο των πραγματικά χρήσιμων δεδομένων.
- Μετάδοση δεδομένων σε υψηλές ταχύτητες.
- Μετατροπή των δεδομένων μεταξύ διαφορετικών μοντέλων και προτύπων.
- Διάγνωση αλλαγών στα δεδομένα από τις πηγές και μεταφορά των νέων δεδομένων.
- Εισαγωγή των δεδομένων στην Αποθήκη Δεδομένων.
- Δημιουργία αντιγράφων τμημάτων των πηγών στην Αποθήκη Δεδομένων.

- Ανάλυση των μεταφερόμενων δεδομένων για τη διάγνωση μη ορθής πληροφορίας.
- Έλεγχος πληρότητας δεδομένων.

Τέλος, η Ε*νημέρωση (Refresh)* της Αποθήκης Δεδομένων είναι η διαδικασία που μεταφέρει τις αλλαγές που συμβαίνουν στα δεδομένα των πηγών εκτελώντας τις αντίστοιχες αλλαγές στα δεδομένα της Αποθήκης. Συνήθως, οι Αποθήκες Δεδομένων ενημερώνονται περιοδικά (π.χ. ανά εβδομάδα, μήνα κλπ.). Υπάρχουν όμως και περιπτώσεις που για τις ανάγκες της ανάλυσης απαιτείται άμεση πρόσβαση σε τρέχοντα δεδομένα, με αποτέλεσμα να επιβάλλεται και η άμεση ενημέρωση των Αποθηκών για κάθε μεταβολή στις πηγές. Ωστόσο, επειδή οι Αποθήκες Δεδομένων συσσωρεύουν μεγάλη ποσότητα δεδομένων η εφαρμογή της διαδικασίας ενημέρωσης καθίσταται απαγορευτική. Γι' αυτό και είναι αναγκαία η διάγνωση των μεταβολών που συμβαίνουν στις πηγές (εισαγωγές, διαγραφές και τροποποιήσεις εγγραφών), ώστε σε κάθε διαδικασία ενημέρωσης να μη γίνονται περιττές διαγραφές και εισαγωγές δεδομένων που στην πραγματικότητα παραμένουν αναλλοίωτα. Πάντως, κάθε φορά η πολιτική ενημέρωσης καθορίζεται από το διαχειριστή της Αποθήκης Δεδομένων βάσει των αναγκών των εφαρμογών ανάλυσης, τη διαθεσιμότητα των πηγών και την κατάσταση του δικτύου που συνδέει την Αποθήκη με τις πηγές.

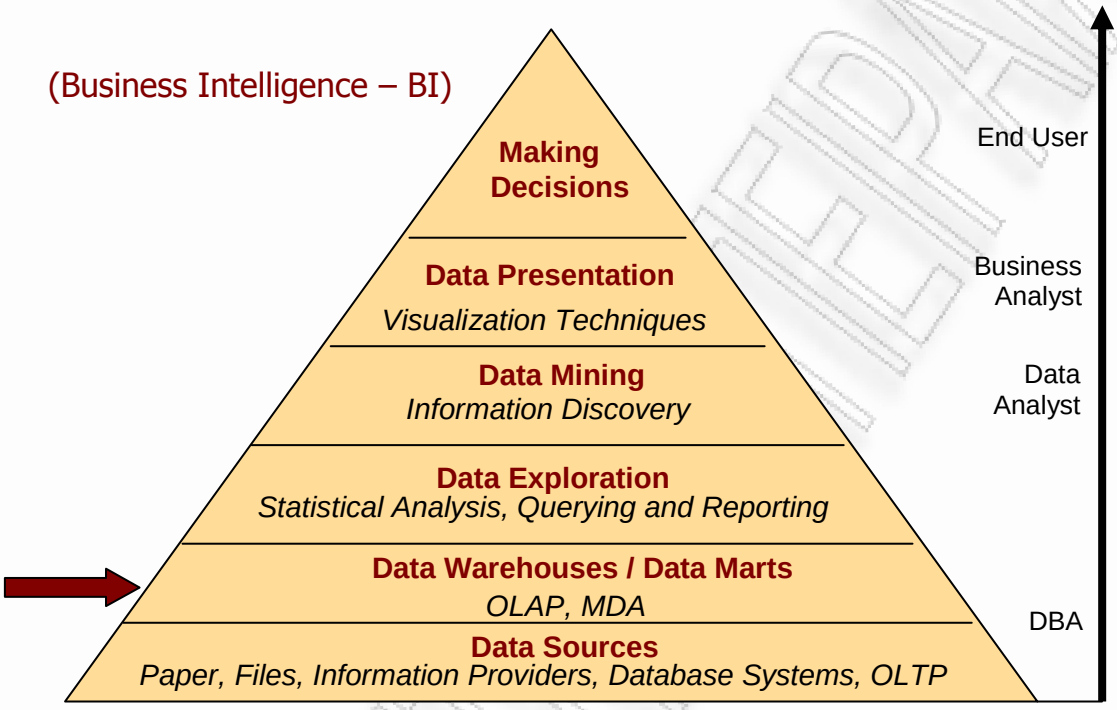
*Αποθήκη Δεδομένων (Data Warehouse), Συλλογές Δεδομένων (Data Marts***)**: Είναι τα συστήματα που αποθηκεύονται τα δεδομένα που παρέχονται προς τους χρήστες, τα οποία υλοποιούνται με τη χρήση Σχεσιακών Συστημάτων Διαχείρισης Βάσεων Δεδομένων. Τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων και η πρόσβαση σε αυτά γίνεται μέσω μιας γλώσσας διαχείρισης δεδομένων που είναι επέκταση της SQL. Εναλλακτικά χρησιμοποιούνται και Πολυδιάστατα Συστήματα Αναλυτικής Επεξεργασίας (Multidimensional OLAP servers), που αποθηκεύουν και διαχειρίζονται δεδομένα με πολυδιάστατο τρόπο. Το κυριότερο πλεονέκτημα των πολυδιάστατων συστημάτων σε σύγκριση με την ευελιξία που χαρακτηρίζει τα Σχεσιακά Συστήματα Διαχείρισης Βάσεων Δεδομένων είναι η δυνατότητά τους να διαχειρίζονται δεδομένα, τα οποία είναι δομημένα με τρόπο που βρίσκεται πιο κοντά στις ανάγκες των εφαρμογών ανάλυσης (OLAP). Οι Συλλογές Δεδομένων περιέχουν τμήματα των δεδομένων της Αποθήκης Δεδομένων και η ύπαρξή τους είναι επιλογή του διαχειριστή του συστήματος. Ο καταμερισμός των δεδομένων της Αποθήκης σε επιμέρους Συλλογές ανά αντικείμενο ή τμήμα γίνεται κυρίως με οργανωτικά κριτήρια και έχει ως στόχο την άμεση και αποδοτική πρόσβαση των εφαρμογών ανάλυσης στα δεδομένα της Αποθήκης.

Βάση Μετα-Δεδομένων (Metadata Repository): Είναι το υποσύστημα αποθήκευσης πληροφορίας σχετικά με τη δομή και λειτουργία όλου του συστήματος και όπως φαίνεται στο Σχήμα 2-2 από το υποσύστημα αυτό υπάρχει πρόσβαση σε όλα τα δομικά στοιχεία της αρχιτεκτονικής της Αποθήκης Δεδομένων. Η κατανόηση και η καταγραφή του περιεχομένου των δεδομένων και της οργάνωσης τους είναι απαραίτητη για την αποδοτική λειτουργία και διαχείριση της Αποθήκης. Τα μετα-δεδομένα πρέπει να περιέχουν:

- Λεξικό δεδομένων (Data Dictionary) που περιέχει τον ορισμό και την περιγραφή των δεδομένων που αποθηκεύονται στην Αποθήκη Δεδομένων και τις μεταξύ τους συσχετίσεις.
- Περιγραφή της ροής των δεδομένων μέσα στο σύστημα.
- Περιγραφή των κανόνων μετατροπής των δεδομένων κατά τη μεταφορά τους.
- Δεδομένα ελέγχου των διαφόρων εκδοχών (versions) των δεδομένων.
- Στατιστικά χρήσης των δεδομένων.
- Πληροφορία σχετικά με τους κανόνες ελέγχου πρόσβασης στην Αποθήκη Δεδομένων.
- Διάφορα ψευδώνυμα (aliases).

Οι Αποθήκες Δεδομένων συγκεντρώνουν μεγάλο όγκο ετερογενών δεδομένων σε πολυδιάστατο χώρο. Η αρχιτεκτονική τους περιλαμβάνει μεταξύ άλλων τον καθαρισμό, την ολοκλήρωση, τη μετατροπή και την εισαγωγή των δεδομένων από τις πηγές στην Αποθήκη. Η σχεδίαση της αρχιτεκτονικής ενός συστήματος Αποθήκης Δεδομένων αποτελεί μια πολύπλοκη διαδικασία και μπορεί να θεωρηθεί ως το κύριο βήμα προεπεξεργασίας των δεδομένων για την εξόρυξη γνώσης που κατ' επέκταση αποτελεί το στάδιο που προηγείται της ολοκλήρωσης της διαδικασίας ανακάλυψης γνώσης. Ουσιαστικά, πάνω στην αρχιτεκτονική της Αποθήκης Δεδομένων βασίζονται οι εφαρμογές ανάλυσης, όπως για παράδειγμα, οι εφαρμογές παραγωγής αναφορών (Reporting), η αναλυτική επεξεργασία δεδομένων με σύνθετα ερωτήματα (OLAP Querying), η ανάλυση για λήψη αποφάσεων (Analysis) και η Εξόρυξη

Γνώσης (Data Mining). Όπως φαίνεται και στο Σχήμα 2-3, οι Αποθήκες Δεδομένων βρίσκονται στη βάση της «Πυραμίδας» της Επιχειρηματικής Ευφυΐας που μπορεί αποκομίσει ο τελικός χρήστης (End User) από πρωτογενή δεδομένα με τη βοήθεια της σύγχρονης τεχνολογίας.

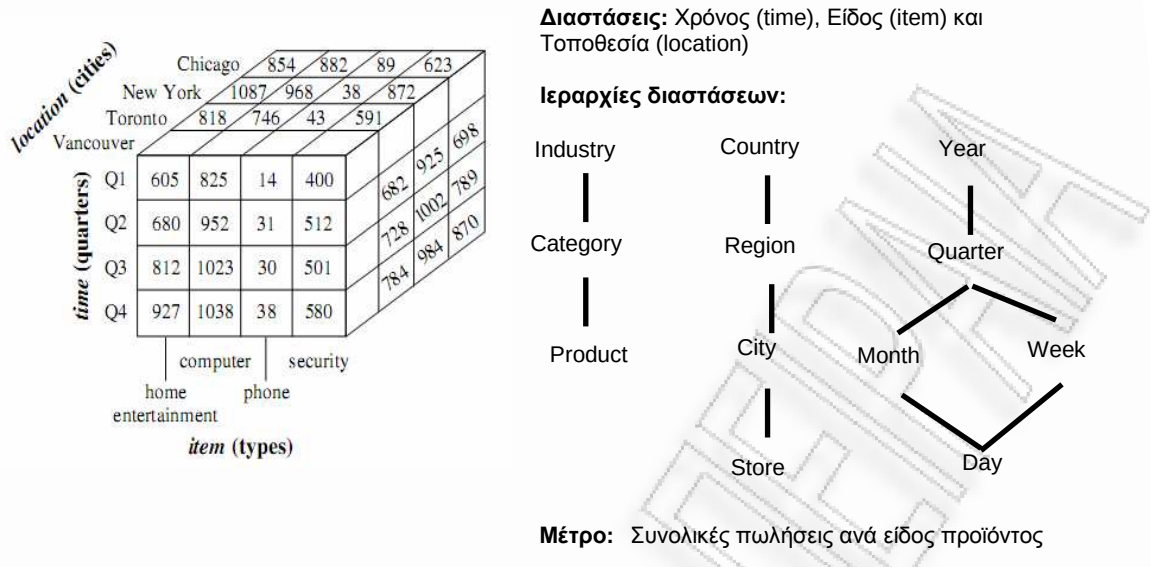


Σχήμα 2-3. «Πυραμίδα» Επιχειρηματικής Ευφυΐας

2.1.3 Η Εννοιολογική Σχεδίαση της Αποθήκης Δεδομένων

Οι Αποθήκες Δεδομένων χρησιμοποιούνται για την αναλυτική επεξεργασία μεγάλου όγκου δεδομένων με σκοπό την απάντηση πολύπλοκων ερωτήσεων σε αποδεκτούς χρόνους. Αυτός είναι και ο λόγος που τόσο η σχεδίαση όσο και η οργάνωση των δεδομένων τους είναι διαφορετική από τις παραδοσιακές σχεσιακές βάσεις δεδομένων. Τα διαγράμματα Οντοτήτων – Συσχετίσεων (Entity – Relationship) και οι τεχνικές κανονικοποίησης των OLTP συστημάτων αποδεικνύονται ακατάλληλα για τη σχεδίαση των Αποθηκών Δεδομένων. Η τεχνική που χρησιμοποιείται στις Αποθήκες Δεδομένων είναι το Μοντέλο Διαστάσεων (Dimensional Modeling) ή διαφορετικά το πολυδιάστατο μοντέλο.

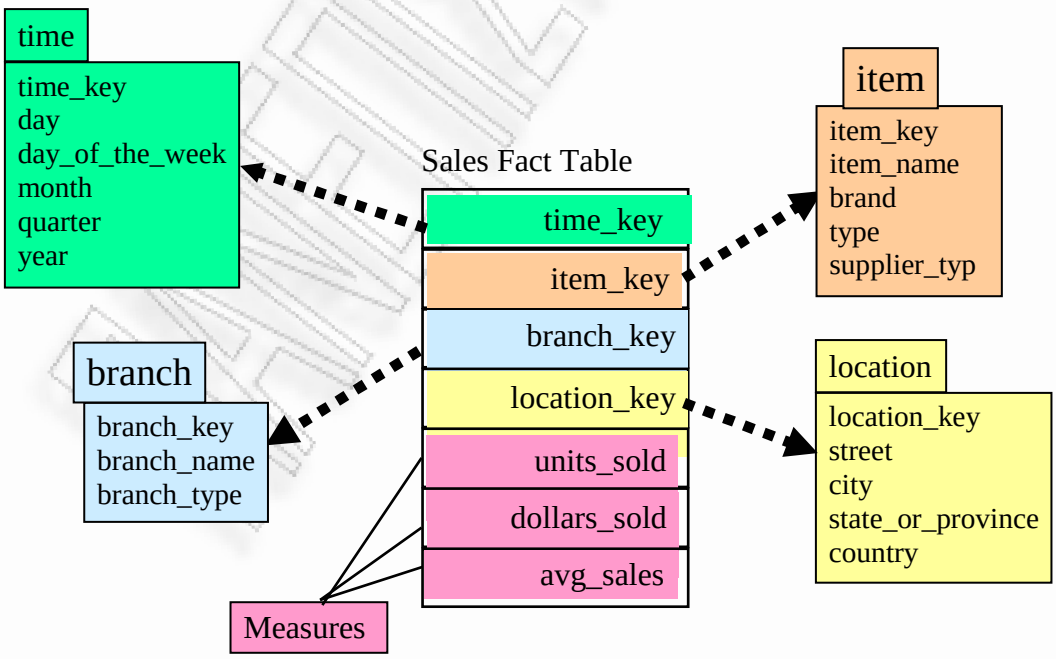
Το Μοντέλο βασίζεται στη θεώρηση των δεδομένων μέσω ενός πολυδιάστατου μοντέλου δεδομένων το οποίο απεικονίζει τα δεδομένα σε μορφή κύβου. Αν και ο όρος κύβος περιγράφει μια γεωμετρική δομή τριών διαστάσεων, στις Αποθήκες Δεδομένων ο κύβος δεδομένων είναι v - διαστάσεων. Δηλαδή, η χρήση ενός κύβου δεδομένων επιτρέπει τη θεώρηση των δεδομένων σε πολλαπλές διαστάσεις. Τα βασικά στοιχεία του Μοντέλου είναι οι *Πίνακες Διαστάσεων (Dimension Tables)* με πληροφορία για τις διαστάσεις του κύβου και οι *Πίνακες Γεγονότων (Fact Tables)* με μέτρα (κάποια μετρήσιμα μεγέθη) και κλειδιά προς τους σχετιζόμενους πίνακες διαστάσεων. Ένα πολυδιάστατο μοντέλο δεδομένων οργανώνεται γύρω από ένα κεντρικό θέμα, όπως για παράδειγμα οι πωλήσεις, το οποίο εμφανίζεται στον πίνακα γεγονότων. Το Σχήμα 2-4 αποτελεί μια απεικόνιση του πολυδιάστατου μοντέλου δεδομένων.



Σχήμα 2-4. Απεικόνιση ενός 3-διάστατου κύβου δεδομένων με διαστάσεις χρόνου, είδους προϊόντος και τοποθεσίας. Το μέτρο συνολικές πωλήσεις προϊόντος εκφράζεται σε χιλιάδες δολάρια.

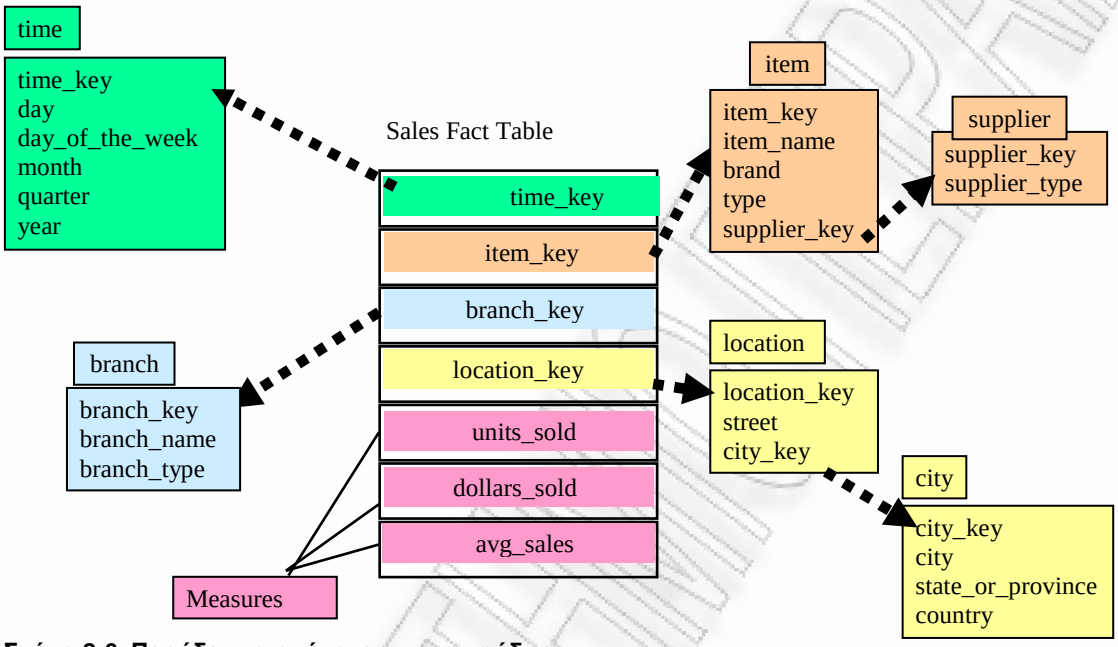
Υπάρχουν δύο βασικές κατηγορίες σχημάτων για τη σχεδίαση των βάσεων των Αποθηκών Δεδομένων, το *Σχήμα Αστέρα (Star Schema)* και το *Σχήμα Χιονονιφάδας (Snowflake Schema)* ενώ μια επιπλέον τεχνική σχεδίασης αποτελούν οι *Αστερισμοί Γεγονότων (Fact Constellations)* ή εναλλακτικά Σχήμα Γαλαξία καθώς πρόκειται για συλλογή σχημάτων αστέρων.

Σχήμα Αστέρα: Αποτελείται από ένα κεντρικό πίνακα γεγονότων και κάποιους από-κανονικοποιημένους πίνακες διαστάσεων. Τα μέτρα είναι τα ενδιαφέροντα μεγέθη υπό μέτρηση (π.χ.units_sold, dollars_sold στον πίνακα SALES). Για κάθε διάσταση του μοντέλου, εισάγεται και ένας πίνακας (π.χ.Time, Branch, Location και Item), ο οποίος περιέχει όλα τα επίπεδα συνάθροισης (levels of aggregation) καθώς τις σχετικές τους ιδιότητες. Το Σχήμα 2-5 παρουσιάζει ένα παράδειγμα σχήματος αστέρα.



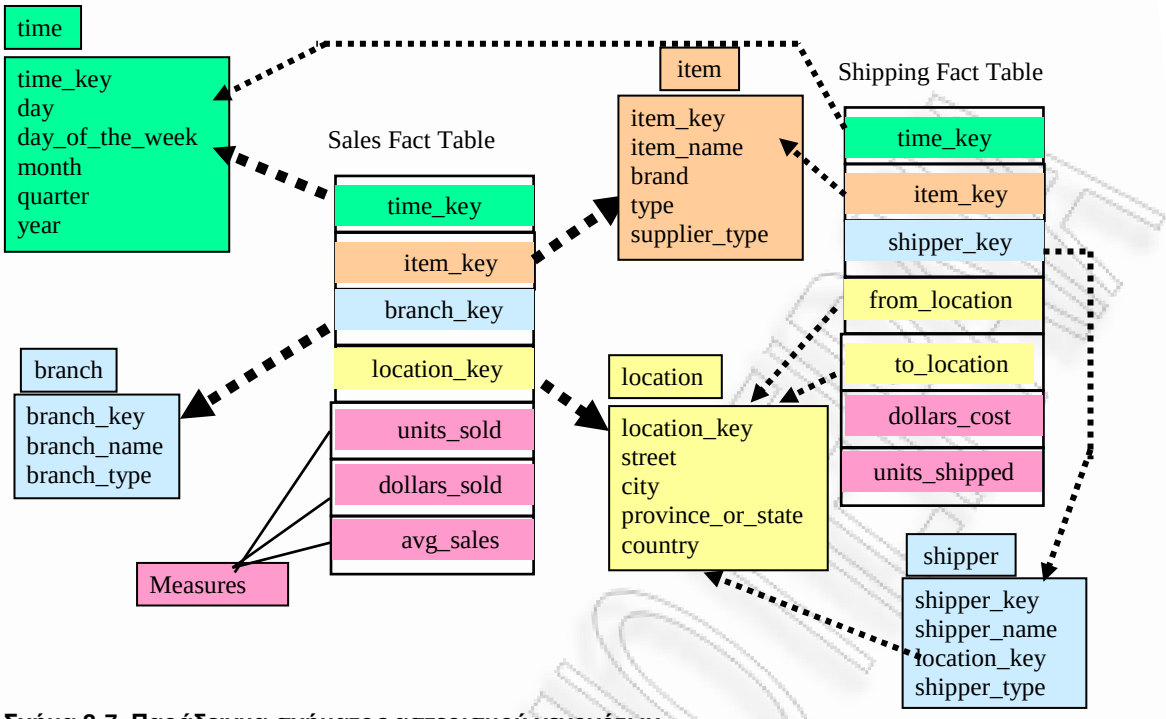
Σχήμα 2-5. Παράδειγμα σχήματος αστέρα.

Σχήμα Χιονονιφάδας: Αποτελεί την κανονικοποιημένη εκδοχή του σχήματος αστέρα. Για κάθε επίπεδο της ιεραρχίας των διαστάσεων εισάγεται και ένας πίνακας. Το σχήμα αυτό εγγυάται την ακεραιότητα των δεδομένων (όπως όλα τα κανονικοποιημένα σχήματα), αλλά είναι πιο αργό στις απαντήσεις των ερωτήσεων. Στο Σχήμα 2-6 παρουσιάζεται ένα παράδειγμα βάσης που έχει το ίδιο περιεχόμενο με τη βάση του Σχήματος 2-5, μόνο που είναι οργανωμένη με το σχήμα χιονονιφάδας. Ουσιαστικά αποτελεί μια βελτίωση του σχήματος αστέρα, όπου η ιεραρχία των διαστάσεων αναπαριστάται με κανονικοποίηση των πινάκων διαστάσεων.



Σχήμα 2-6. Παράδειγμα σχήματος χιονονιφάδας.

Αστερισμός Γεγονότων: Το σχήμα αυτό χρησιμοποιείται όταν χρειάζεται να υπάρχουν πολλοί πίνακες γεγονότων, οι οποίοι να μοιράζονται τους πίνακες διαστάσεων. Είναι συχνό φαινόμενο στις Αποθήκες Δεδομένων αλλά πιο σπάνιο στη σχεδίαση Συλλογών Δεδομένων. Το Σχήμα 2-7 απεικονίζει ένα παράδειγμα Αστερισμού Γεγονότων με χρήση του ίδιου περιεχομένου βάσης, όπως και στα Σχήματα 2-5, 2-6.



Σχήμα 2-7. Παράδειγμα σχήματος αστερισμού γεγονότων.

Η αναλυτική επεξεργασία δεδομένων είναι τμήμα των εφαρμογών στήριξης αποφάσεων και των στρατηγικών πληροφοριακών συστημάτων. Η λειτουργία της αναλυτικής επεξεργασίας δεδομένων (OLAP) χαρακτηρίζεται από τη δυναμική πολυδιάστατη ανάλυση των δεδομένων ενός οργανισμού με εκτέλεση ερωτήσεων πάνω στα δεδομένα. Οι ερωτήσεις έχουν συγκεκριμένη και πολύπλοκη δομή, ενώ η πληροφορία που αντλούν έχει πολυδιάστατο χαρακτήρα. Τα πολυδιάστατα μοντέλα δεδομένων περιέχουν ν-διάστατους πίνακες που συχνά αποκαλούνται υπερκύβοι (cubes ή hypercubes). Κάθε διάσταση έχει μία ιεραρχία επιπέδων, π.χ. η διάσταση "Γεωγραφική τοποθεσία" έχει τα επίπεδα πόλη, περιοχή, χώρα. Οι τιμές (μετρήσιμα μεγέθη) που περιέχουν οι υπερκύβοι αντιστοιχούν στις στήλες των σχεσιακών πινάκων. Ένα παράδειγμα αναλυτικής επεξεργασίας δεδομένων είναι μια εφαρμογή που εκτελεί ερωτήσεις για να μπορεί να έχει συγκεντρωτικά δεδομένα για τις πωλήσεις ανά προϊόν, ανά μήνα και ανά περιοχή ενός οργανισμού. Η παρουσίαση των αποτελεσμάτων των πωλήσεων μπορεί να προκαλέσει το χρήστη στην εκτέλεση μιας πιο συγκεντρωτικής ερώτησης, ώστε να πάρει ως απάντηση τα δεδομένα που αφορούν τις ετήσιες πωλήσεις ανά προϊόν και περιοχή, ή να εκτελέσει μια πιο λεπτομερή ερώτηση παίρνοντας ως απάντηση τις μηνιαίες πωλήσεις κάθε προϊόντος ανά συγκεκριμένο πελάτη.

Οι κύβοι δίνουν τη δυνατότητα πλοήγησης στις ιεραρχίες των διαστάσεών τους. Η πλοήγηση είναι δυνατή από τις λειτουργίες τις οποίες παρέχουν. Οι OLAP λειτουργίες που γίνονται συνήθως στους κύβους είναι οι παρακάτω ενώ σχετικά παραδείγματα παρουσιάζονται στο Σχήμα 2-8:

Συσσώρευση (Roll-up): Πρόκειται για μια λειτουργία με την οποία εκτελείται ένα βήμα ανόδου στην ιεραρχία μιας διάστασης (π.χ. από ημέρα σε μήνα). Ο κύβος που προκύπτει από τη λειτουργία της συνάθροισης της πληροφορίας περιέχει πιο ομαδοποιημένα δεδομένα, με βάση τη διάσταση στην οποία έγινε η ομαδοποίηση. Η ανάβαση στην ιεραρχία μπορεί να συνεχιστεί με όμοιο τρόπο.

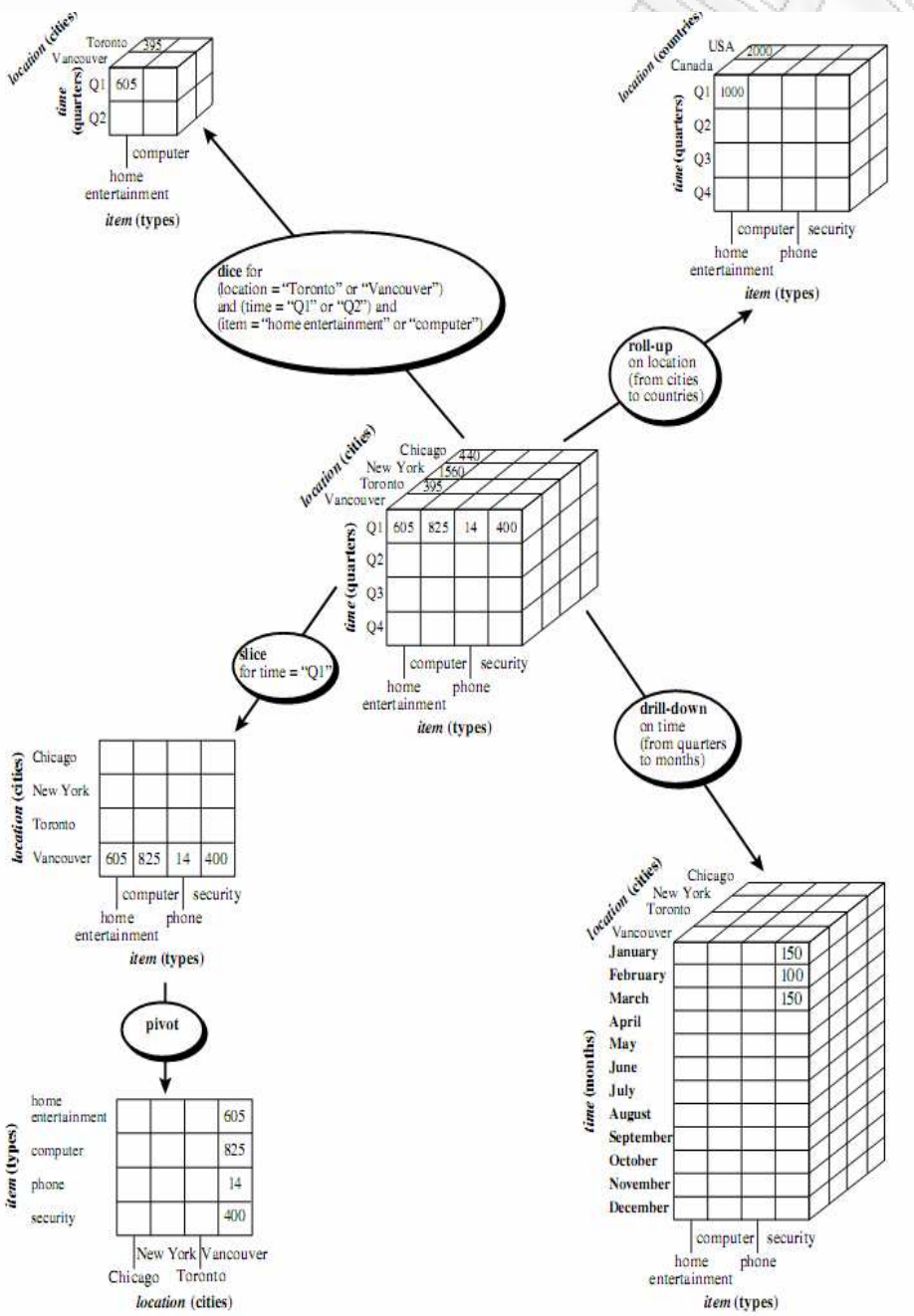
Εμβάθυνση (Drill-down): Είναι η αντίστροφη πράξη του roll-up, όπου εκτελείται ένα βήμα καθόδου από ένα υψηλότερο επίπεδο της ιεραρχίας μιας διάστασης σε ένα χαμηλότερο.

Τεμαχισμός (Slice): Πρόκειται για λειτουργία επιλογής δεδομένων σε μία συγκεκριμένη διάσταση. Ένα επίπεδο (slice) είναι ένα υποσύνολο ενός υπερκύβου σύμφωνα με μια περιοχή τιμών ή μια συγκεκριμένη τιμή ενός επιπέδου διάστασης (οριζόντιος τεμαχισμός).

Κομμάτιασμα (Dice): Πρόκειται για λειτουργία επιλογής δεδομένων από δύο ή και περισσότερες διαστάσεις (κάθετος τεμαχισμός).

Περιστροφή (Pivot): Πρόκειται για λειτουργία αλλαγής της διάταξης των διαστάσεων ώστε να διευκολυνθεί η ανάλυση. Κατά την περιστροφή, δεν μεταβάλλονται ούτε μειώνονται τα δεδομένα του υπερκύβου, απλά αλλάζει ο τρόπος παρουσίασής τους στην εφαρμογή ανάλυσης.

Άλλες OLAP λειτουργίες μπορεί να περιλαμβάνουν την κατάταξη των ν-πρώτων ή των ν-τελευταίων αντικειμένων σε λίστες, τον υπολογισμό κινητών μέσων, ρυθμών ανάπτυξης, κερδών, τόκων, μετατροπές νομισμάτων και στατιστικές συναρτήσεις. Επίσης, η αναλυτική επεξεργασία δεδομένων υποστηρίζει λειτουργικά μοντέλα για πρόβλεψη, ανάλυση τάσεων της αγοράς και στατιστική ανάλυση. Με αυτό το περιεχόμενο, η μηχανή OLAP αποτελεί τελικά ένα πολύ δυνατό εργαλείο ανάλυσης δεδομένων.



Σχήμα 2-8 . Παραδείγματα OLAP λειτουργιών σε πολυδιάστατα δεδομένα.

2.2 Εξόρυξη Γνώσης από Δεδομένα

Τα τελευταία χρόνια, κυρίως λόγω των δυνατοτήτων που προσφέρουν οι νέες τεχνολογικές εξελίξεις, τεράστιος είναι ο όγκος των δεδομένων κάθε είδους που αποθηκεύονται σε αρχεία και βάσεις δεδομένων. Εκείνοι οι οποίοι έχουν την ικανότητα να συλλέγουν πληροφορίες και δεδομένα, και έπειτα να τα αναλύουν και να τα αξιοποιούν, μοιραία είναι σε θέση να πρωταγωνιστήσουν σε όποιο πεδίο δραστηριοποιούνται. Η πληροφορία και η αξιοποίησή της, καθώς και η ανάλυση διάφορων δεδομένων τα οποία μπορούν να συλλεχθούν δίνουν τη δυνατότητα σε κάθε ενδιαφερόμενο να αποκτήσει ένα ανταγωνιστικό πλεονέκτημα στο χώρο στον οποίο δραστηριοποιείται και να πάρει τελικά τις βέλτιστες αποφάσεις σε θέματα που τον αφορούν. Τέτοιου είδους αναλύσεις, που λαμβάνουν χώρα σε ποιοτικά αλλά και αριθμητικά δεδομένα γίνονται με τη βοήθεια τεχνικών εξόρυξης γνώσης από δεδομένα (Data Mining), οι οποίες παρέχουν τη δυνατότητα εξαγωγής κανόνων και άρα αποφάσεων με τη βοήθεια των ηλεκτρονικών υπολογιστών.

Η σημερινή εξέλιξη στις λειτουργίες και στα προϊόντα της εξόρυξης γνώσης είναι αποτέλεσμα της πολυετούς επιρροής διάφορων επιστημονικών κλάδων όπως, της Μηχανικής Μάθησης (Machine Learning), της Αναγνώρισης Κανόνων (Pattern Recognition), των Βάσεων Δεδομένων (Databases), της Στατιστικής (Statistics), της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI) και των Έμπειρων Συστημάτων (Expert Systems). Οι περισσότεροι αλγόριθμοι και τεχνικές προέρχονται από αυτά τα πεδία. Η βάση όλων των παραπάνω είναι η απόσπαση κανόνων που περιέχουν γνώση, μέσα από πλήθος δεδομένων.

Στη συνέχεια θα γίνει αναφορά στο τι ακριβώς αντιπροσωπεύει η διαδικασία εξόρυξης γνώσης από δεδομένα και τι είδους δεδομένα χρησιμοποιεί, στα στάδια της εξόρυξης γνώσης από βάσεις δεδομένων ή πολυδιάστατα μοντέλα, στις τεχνικές εξόρυξης γνώσης που χρησιμοποιούνται γενικά, καθώς και στις νέες τάσεις που επικρατούν στην αναπτυσσόμενη τεχνολογία γύρω από το επιστημονικό πεδίο της εξόρυξης γνώσης από δεδομένα.

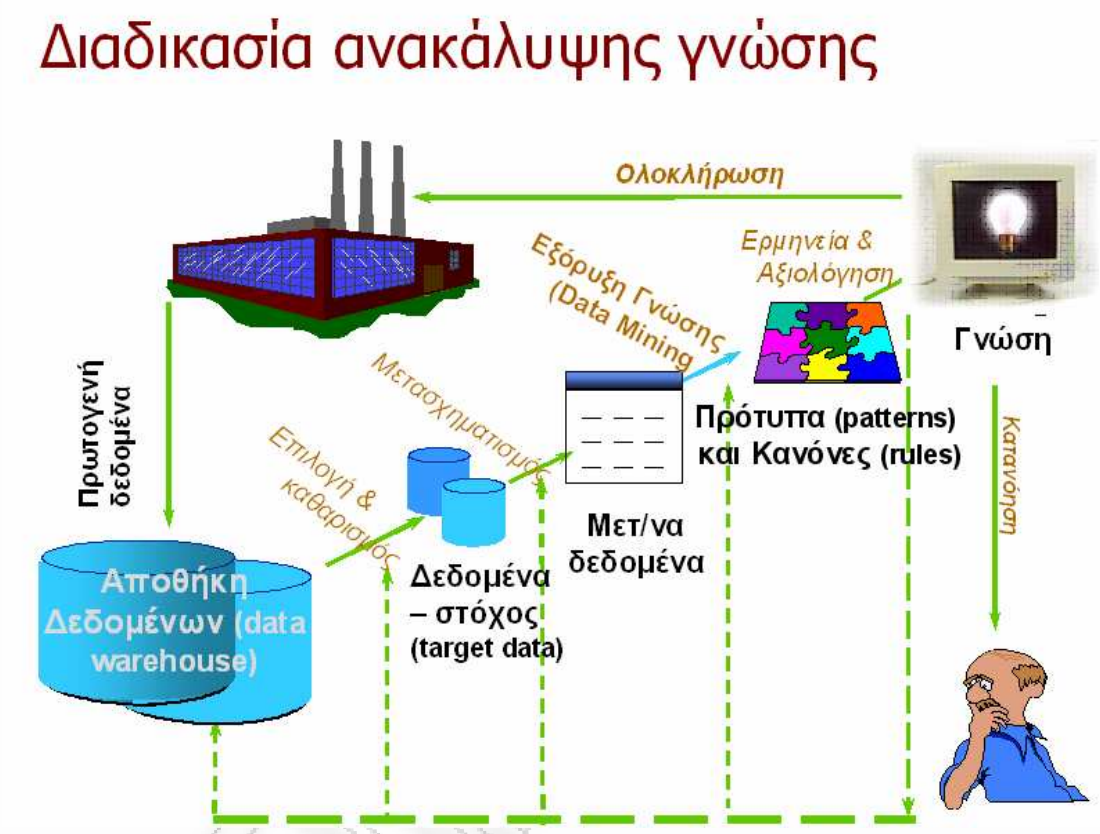
2.2.1 Εξόρυξη Γνώσης από Δεδομένα και Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων

Ο τεράστιος όγκος δεδομένων που αποθηκεύεται σε αρχεία, βάσεις δεδομένων και άλλα αποθηκευτικά μέσα επιβάλλει την ανάπτυξη δυναμικών μέσων ανάλυσης και ερμηνείας τέτοιων δεδομένων με σκοπό την εξαγωγή χρήσιμης γνώσης και τη λήψη αποφάσεων. Η διαδικασία Data Mining, η ελληνική απόδοση της οποίας είναι “Εξόρυξη γνώσης από Δεδομένα ή Ανεύρεση Γνώσης από Δεδομένα”, είναι η αναλυτική διαδικασία η οποία έχει σχεδιαστεί με σκοπό να αναλύει και να εξερευνεί δεδομένα σε μεγάλες ποσότητες και στη συνέχεια να δημιουργεί κανόνες και σχέσεις μεταξύ των μεταβλητών που ενδιαφέρουν να ερευνηθούν. Η όλη διαδικασία βασίζεται στη χρησιμοποίηση αλγορίθμων που αναζητούν κανόνες μεταξύ των μεταβλητών των δεδομένων και έπειτα βρίσκουν συσχετισμούς ή κανόνες μέσα από τεράστιες βάσεις αποθηκευμένων δεδομένων / πληροφοριών. Επίσης η διαδικασία εξόρυξης γνώσης από δεδομένα αναφέρεται συχνά και ως Πληροφοριακή Τεχνολογία (Computerized Technology) η οποία χρησιμοποιεί πολύπλοκους αλγόριθμους που δημιουργούν κανόνες και σχέσεις αναλύοντας τεράστιες βάσεις δεδομένων, με στόχο τη λήψη στρατηγικών αποφάσεων. Η εξόρυξη γνώσης από τα δεδομένα μπορεί να οριστεί απλά ως η εύρεση πληροφορίας που είναι κρυμμένη σε μεγάλες ποσότητες δεδομένων. Υπάρχουν πολλοί άλλοι όροι που έχουν παρόμοια σημασία με την εξόρυξη γνώσης από δεδομένα όπως, η εξαγωγή γνώσης (knowledge extraction), η ανάλυση δεδομένων/προτύπων (data/pattern analysis), η αρχαιολογία δεδομένων (data archaeology) και η εκβάθυνση δεδομένων (data dredging).

Οι όροι ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases – KDD) και εξόρυξη γνώσης από δεδομένα συχνά αναφέρονται στην ίδια έννοια, δηλαδή στη διαδικασία ανακάλυψης χρήσιμων, συνήθως κρυμμένων προτύπων από τα δεδομένα. Γενικά, έχει οριστεί ότι “Η ανακάλυψη γνώσης είναι η μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, πρωτότυπων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων από τα δεδομένα” [4]. Επίσης, ένας δεύτερος ορισμός που δίνεται είναι ότι η διαδικασία ανακάλυψης γνώσης περιλαμβάνει “Το σχεδιασμό αποθηκών δεδομένων (data warehousing), τη συλλογή δεδομένων στόχου, τον καθαρισμό, την προεπεξεργασία, τη μετατροπή και την ελαχιστοποίηση των δεδομένων, την εξόρυξη γνώσης, την επιλογή μοντέλου ή συνδυασμού μοντέλων, την

αξιολόγηση και τελικά την ενοποίηση και χρησιμοποίηση της εξαγόμενης γνώσης" όπως φαίνεται στο Σχήμα 2-9 [3].

Γενικά, η ανακάλυψη γνώσης σε βάσεις δεδομένων είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων από τα δεδομένα ενώ η εξόρυξη γνώσης από δεδομένα είναι η χρήση αλγόριθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με τη διαδικασία KDD. Η διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων είναι μια επαναληπτική διαδικασία και που εκτός από την εξόρυξη γνώσης, περιλαμβάνει μια μεθοδολογία για την εξαγωγή και την προετοιμασία της γνώσης, καθώς επίσης και τη λήψη αποφάσεων σχετικών με τις ενέργειες που πρέπει να γίνουν όταν ολοκληρωθεί η εξόρυξη γνώσης.



Σχήμα 2-9: Η Εξόρυξη Γνώσης αποτελεί τον πυρήνα της διαδικασίας ανακάλυψης γνώσης σε βάσεις δεδομένων.

Η διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων είναι μια διαλογική και επαναληπτική διαδικασία, δηλαδή μπορεί να απαιτηθεί η επιστροφή σε κάποιο προηγούμενο βήμα όπως φαίνεται στο παραπάνω σχήμα. Η διαδικασία KDD μπορεί να διαχωριστεί στα παρακάτω βήματα:

Ορισμός του προβλήματος (Defining the problem): Στο βήμα αυτό ορίζεται το πλαίσιο δράσης της διαδικασίας KDD, δηλαδή καθορίζονται οι προσδοκίες για τα αποτελέσματα από την εξόρυξη γνώσης που περιλαμβάνουν ουσιαστικά τις απαιτήσεις των αναλυτών, τις στρατηγικές marketing, τις προβλέψεις και την υποστήριξη αποφάσεων.

Συλλογή Δεδομένων (Data Collection): Το βήμα αυτό περιλαμβάνει τον εντοπισμό των δεδομένων που είναι διαθέσιμα, την απόκτηση επιπρόσθετων δεδομένων που είναι αναγκαία για την ανάλυση και τελικά την ενσωμάτωση όλων αυτών σε ένα σύνολο δεδομένων το οποίο θα περιλαμβάνει τα χαρακτηριστικά (attributes) που θα ληφθούν υπόψη. Οι αλγόριθμοι εξόρυξης γνώσης εκπαιδεύονται και ανακαλύπτουν πρότυπα από τα δεδομένα που είναι κάθε

φορά διαθέσιμα, και επομένως σε κάθε περίπτωση είναι απαραίτητη η μέγιστη δυνατή συλλογή χαρακτηριστικών.

Καθαρισμός των δεδομένων (Data Cleaning): Η αξιοπιστία των δεδομένων αποτελεί ένα πολύ σημαντικό σημείο στη διαδικασία KDD. Στο βήμα αυτό πραγματοποιείται καθαρισμός των δεδομένων, δηλαδή διαχείριση των ελλιπών τιμών (missing values) και απομάκρυνση δεδομένων με θόρυβο ή δεδομένων με ακραίες τιμές (outliers). Ο καθαρισμός των δεδομένων επιτυγχάνεται με τη χρησιμοποίηση σύνθετων στατιστικών μεθόδων ή αλγόριθμων εξόρυξης γνώσης (π.χ. Bayesian formula ή δέντρα απόφασης).

Μετασχηματισμός των δεδομένων (Data Transformation): Στο βήμα αυτό τα δεδομένα μετασχηματίζονται σε κατάλληλες μορφές για εξόρυξη γνώσης. Αυτό επιτυγχάνεται με την εφαρμογή μεθόδων εξομάλυνσης, κανονικοποίησης των τιμών των χαρακτηριστικών και διακριτοποίησης των συνεχών μεταβλητών καθώς κάποιοι αλγόριθμοι εξόρυξης γνώσης συμπεριφέρονται καλύτερα όταν χρησιμοποιούνται διακριτά δεδομένα.

Επιλογή μεθόδου εξόρυξης γνώσης από τα δεδομένα (Data Mining): Σε αυτό το βήμα εφαρμόζονται έξυπνες τεχνικές (κατηγοριοποίηση - classification, συσταδοποίηση - clustering, κανόνες συσχετίσεων - association rules στις οποίες θα γίνει εκτενής αναφορά σε επόμενες παραγράφους) εξαγωγής δυνητικά χρήσιμων προτύπων από τα δεδομένα. Οι δύο βασικοί στόχοι της εξόρυξης γνώσης είναι η περιγραφή και η πρόβλεψη. Εφόσον έχει οριστεί η στρατηγική που θα ακολουθηθεί, επιλέγεται και εκτελείται ο αλγόριθμος εξόρυξης γνώσης από δεδομένα. Η απόδοση και τα εξαγόμενα αποτελέσματα εξαρτώνται από τα προηγούμενα βήματα.

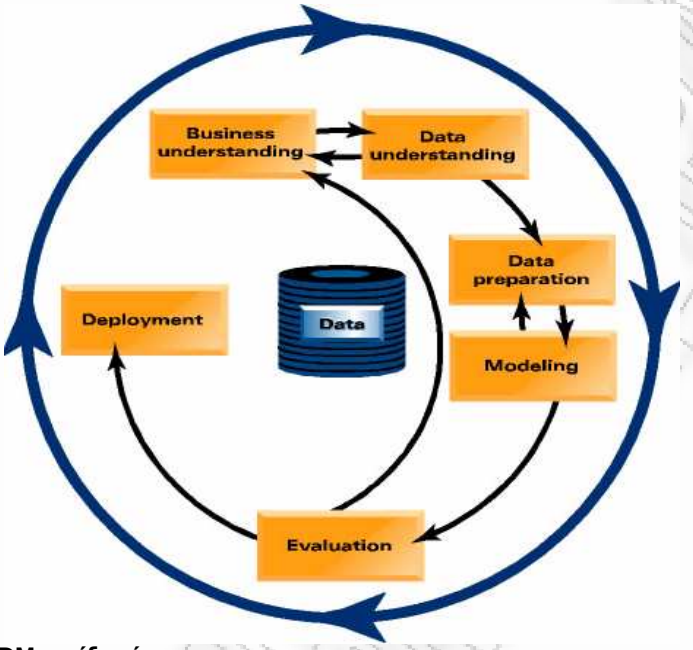
Αξιολόγηση προτύπου (Pattern Evaluation): Σε αυτό το βήμα γίνεται εκτίμηση και ερμηνεία των εξορυχθέντων προτύπων (κανόνες, συσχετισμοί, αξιοπιστία κλπ) σε σύγκριση με τους στόχους που είχαν τεθεί κατά τον αρχικό ορισμό του προβλήματος, δηλαδή στο πρώτο βήμα της διαδικασίας. Ουσιαστικά, αξιολογείται η χρησιμότητα του μοντέλου και τεκμηριώνεται η γνώση που ανακαλύφθηκε η οποία είναι διαθέσιμη για περαιτέρω χρήση.

Παρουσίαση της ανακαλυφθείσας γνώσης (Knowledge representation): Σε αυτό το τελευταίο βήμα η γνώση που ανακαλύφθηκε οπτικοποιείται και παρουσιάζεται στο χρήστη. Δηλαδή, χρησιμοποιούνται τεχνικές οπτικοποίησης που βοηθούν τους χρήστες στην κατανόηση και ερμηνεία των αποτελεσμάτων της εξόρυξης γνώσης και κατ' επέκταση την αποτελεσματικότητας της χρήσης της διαδικασίας KDD.

Τα παραπάνω βήματα συνδυάζονται και επαναλαμβάνονται κατά τη διαδικασία KDD, καθώς μετά από κάθε αξιολόγηση και παρουσίαση της ανακαλυφθείσας γνώσης στο χρήστη, μπορεί να γίνει εκ νέου συλλογή δεδομένων και μετασχηματισμός αυτών, εκτέλεση διαφορετικών αλγορίθμων κλπ., ώστε να εξαχθούν καλύτερα και πιο κατάλληλα αποτελέσματα.

2.2.2 Η Μεθοδολογία Εξόρυξης Γνώσης από Δεδομένα Crisp-DM

Το 1996 οι εταιρίες SPSS, NCR και DaimlerChrysler πρότειναν τη μεθοδολογία εξόρυξης γνώσης Crisp-DM (Cross-Industry Standard Process for Data Mining), η οποία δημοσιεύτηκε το 2000 μετά από επιχορήγηση της Ευρωπαϊκής Επιτροπής. Η Crisp-DM δεν περιγράφει μια συγκεκριμένη τεχνική εξόρυξης γνώσης αλλά εστιάζει στον κύκλο ζωής μιας διαδικασίας εξόρυξης γνώσης που όπως απεικονίζεται στο Σχήμα 2-10 αποτελείται έξι φάσεις.



Σχήμα 2-10: CRISP-DM οι έξι φάσεις

Κατανόηση της περιοχής δράσης (Business Understanding Phase): Η πρώτη φάση της τυποποιημένης διαδικασίας Crisp-DM μπορεί να οριστεί ως η φάση της κατανόησης του ερευνητικού πλαισίου. Στη φάση αυτή καθορίζονται οι αντικειμενικοί σκοποί και οι απαιτήσεις της επιχείρησης ή του τομέα έρευνας βάσει των οποίων προσδιορίζεται το πρόβλημα που θα αντιμετωπιστεί με τις τεχνικές της εξόρυξης γνώσης. Ουσιαστικά, προετοιμάζει μια προκαταρκτική στρατηγική για την επίτευξη των αντικειμενικών σκοπών που έχουν τεθεί λαμβάνοντας υπόψη τους σχετικούς τεχνικούς, οικονομικούς, επιχειρηματικούς και οργανωτικούς παράγοντες και αναγνωρίζοντας τυχόν περιορισμούς (π.χ. πηγές δεδομένων, προστασία δεδομένων προσωπικού χαρακτήρα).

Κατανόηση των δεδομένων (Data Understanding Phase): Στη φάση αυτή αφού πρώτα προσδιοριστούν οι πηγές από τις οποίες θα αντληθούν δεδομένα γίνεται συλλογή των δεδομένων. Στη συνέχεια περιγράφεται το περιεχόμενο των δεδομένων και ορίζονται οι απαιτήσεις προπαρασκευής των δεδομένων. Εξερευνούνται οι συσχετίσεις μεταξύ των δεδομένων και αναδεικνύονται θέματα που αφορούν την ποιότητα των δεδομένων (ελλιπείς τιμές, περίεργες κατανομές των τιμών των δεδομένων).

Προετοιμασία των δεδομένων (Data Preparation Phase): Η φάση αυτή περιλαμβάνει την ολοκλήρωση – ομογενοποίηση των δεδομένων τα οποία προέρχονται συνήθως από ετερογενείς πηγές, τη δειγματοληπτική επιλογή υποσυνόλων δεδομένων που θα χρησιμοποιηθούν ως σύνολα εκπαίδευσης και τέλος την αξιολόγηση της ποιότητας των δεδομένων και την εφαρμογή τεχνικών καθαρισμού και μετασχηματισμού τους.

Μοντελοποίηση (The Modelling Phase): Στη φάση αυτή δημιουργείται το μοντέλο που θα χρησιμοποιηθεί κατά τη διαδικασία εξόρυξης γνώσης, δηλαδή επιλέγεται ο κατάλληλος αλγόριθμος που θα εκτελεστεί, γίνεται μελέτη της συμπεριφοράς τους μοντέλου, εξετάζεται η προσαρμογή του μοντέλου, ερευνούνται λάθη που ανακύπτουν και επαναληπτικά προσαρμόζονται οι παραμετρικές ρυθμίσεις για εξαγωγή καλύτερων αποτελεσμάτων.

Αξιολόγηση (The Evaluation Phase): Η αξιολόγηση του μοντέλου γίνεται με εκτίμηση των εξαγόμενων αποτελεσμάτων από τους ενδιαφερόμενους τελικούς χρήστες (π.χ. αναλυτές, διευθυντικά στελέχη επιχείρησης) και στη συνέχεια αποτιμάται η χρησιμότητα των αποτελεσμάτων από την οπτική γωνία της επιχείρησης/οργανισμού (π.χ. προσδιορισμός ομάδων ελέγχου, αναμενόμενο κέρδος από την επένδυση σε τεχνολογία ανάλυσης δεδομένων). Στη συνέχεια γίνεται ανασκόπηση της διαδικασίας και καθορίζονται τα επόμενα βήματα που θα ακολουθηθούν, δηλαδή η προοπτική εγκατάστασης του μοντέλου, η αρχιτεκτονική τους συστήματος και οι παράγοντες – μετρήσιμα μεγέθη που θα βοηθήσουν στην επιτυχή εγκατάσταση του μοντέλου.

Εγκατάσταση (The Deployment Phase): Αποτελεί την τελευταία φάση της μεθοδολογίας Crisp-DM κατά την οποία γίνεται παρουσίαση της εξαγόμενης γνώσης στους ενδιαφερόμενους χρήστες με τεχνικές οπτικοποίησης, παράγονται αναφορές και αποτιμάται η αποτελεσματικότητα και η χρησιμότητα του μοντέλου με την κατάρτιση μιας τελικής μελέτης στην οποία τεκμηριώνεται η όλη διαδικασία.

Η Crisp-DM προήλθε από την ανάγκη τυποποίησης της διαδικασίας εξόρυξης γνώσης η οποία πρέπει να είναι αξιόπιστη και να παρέχει δυνατότητες για επανάληψη κάποιων βημάτων ακόμα και από χρήστες με μικρό θεωρητικό υπόβαθρο πάνω στην ανακάλυψη κρυμμένης πληροφορίας από τα δεδομένα. Γενικά, η Crisp-DM είναι μια μεθοδολογία εξόρυξης γνώσης που προσδιορίζει τα μοντέλα επεξεργασίας των δεδομένων, είναι προσιτή στον καθένα και παρέχει ένα πλήρες και αρκετά λεπτομερές προσχέδιο των απαιτήσεων και των αντικειμενικών σκοπών της διαδικασίας εξόρυξης γνώσης της οποίας ο κύκλος ζωής αποτελείται από έξι φάσεις όπως αναλύθηκαν προηγουμένως. Ουσιαστικά, η παραπάνω μεθοδολογία τυποποιεί τα βήματα της διαδικασίας ανακάλυψης γνώσης σε βάσεις δεδομένων όπως φαίνεται και από τη σύγκριση των έξι φάσεων με τα βήματα της KDD της προηγούμενης παραγράφου.

2.2.3 Πληθώρα Αποθηκευμένης Πληροφορίας – Εξόρυξη Γνώσης από Διαφορετικούς Τύπους Δεδομένων

Τα σύγχρονα αποθηκευτικά μέσα παρέχουν τη δυνατότητα συλλογής μεγάλου όγκου δεδομένων, από απλές αριθμητικές μετρήσεις και έγγραφα κειμένου έως πιο σύνθετη πληροφορία όπως χωρικά δεδομένα, κανάλια πολυμέσων και έγγραφα υπέρ-κειμένου. Στη συνέχεια παρουσιάζεται η ποικιλία της πληροφορίας που συλλέγεται σε ψηφιακή μορφή σε βάσεις δεδομένων και απλά αρχεία με τη μορφή λίστας η οποία βέβαια δεν είναι απαραίτητα αποκλειστική:

Επιχειρηματικές Συναλλαγές (Business transactions): Στην επιχειρηματική βιομηχανία κάθε συναλλαγή που καταγράφεται συνήθως σχετίζεται με το χρόνο και έχει διάφορες μορφές όπως αγορά, ανταλλαγή, τραπεζική συναλλαγή, απόθεμα στις αποθήκες, αποτέλεσμα οικονομικής χρήσης κλπ. Γενικά, χάρη στη διαδεδομένη χρήση του γραμμοκώδικα (bar code) εκατομμύρια συναλλαγών αποθηκεύονται καθημερινά και το πρόβλημα που ανακύπτει είναι η αποτελεσματική χρήση τους σε εύλογο χρονικό διάστημα για τη λήψη αποφάσεων.

Επιστημονικά Δεδομένα (Scientific Data): Διάφορα επιστημονικά ερευνητικά κέντρα συσσωρεύουν τεράστιο όγκο δεδομένων τα οποία χρήζουν ανάλυσης, ωστόσο το πρόβλημα είναι ότι πιο γρήγορα αποθηκεύονται καινούργια δεδομένα απ' ότι μπορούν να αναλυθούν τα ήδη συγκεντρωμένα.

Ιατρικά και Προσωπικά δεδομένα (Medical and personal data): Κυβερνητικοί φορείς, νοσηλευτικά ιδρύματα, εταιρίες και οργανισμοί αποθηκεύουν πολύ σημαντικές ποσότητες προσωπικών δεδομένων με σκοπό τη διαχείριση των ανθρώπινων πόρων, την καλύτερη κατανόηση της αγοράς ή απλώς την εξυπηρέτηση της πελατείας τους. Παρόλο που αυτό το είδος δεδομένων εμπίπτει σε θέματα ιδιωτικότητας και προστασίας των δεδομένων, η πληροφορία αυτή συλλέγεται, χρησιμοποιείται και μοιράζεται καθώς συσχετιζόμενη με άλλα δεδομένα αναδεικνύει θέματα γύρω από τη συμπεριφορά και τις προτιμήσεις των πελατών.

Βίντεο Παρακολούθησης και Φωτογραφίες (Surveillance video and pictures): Η πληροφορία που αποθηκεύεται στα μέσα βιντεοσκόπησης και ψηφιοποιείται για μελλοντική χρήση και ανάλυση.

Δορυφόροι (Satellite sensing): Υπάρχουν αμέτρητοι δορυφόροι (στατικοί ή σε τροχιά) γύρω από την υφήλιο, οι οποίοι καθημερινά συλλέγουν δεδομένα και φωτογραφίες ώστε μελλοντικά να αναλυθούν.

Παιχνίδια (Games): Καθημερινά συλλέγονται τεράστιες ποσότητες δεδομένων και στατιστικών για παιχνίδια, παίχτες και αθλητές. Τα δεδομένα αυτά χρησιμοποιούνται από προπονητές και αθλητές για τη βελτίωση των επιδόσεων και την καλύτερη κατανόηση των αντιπάλων καθώς και από δημοσιογράφους στις ανταποκρίσεις τους.

Ψηφιακά Μέσα (Digital media): Η εξέλιξη των ψηφιακών μέσων αποθήκευσης βοήθησε στην ψηφιοποίηση συλλογών εικόνας και ήχου που διαθέτουν τα μέσα ενημέρωσης με αποτέλεσμα να βελτιώσουν τη διαχείριση των οπτικοακουστικών αποθεμάτων τους.

Δεδομένα Σχεδιαστικών Συστημάτων και Δεδομένα Λογισμικού Εφαρμοσμένης Μηχανικής (CAD – Computer Assisted Systems and Software Engineering Data): Τα συστήματα αυτά παράγουν τεράστιες ποσότητες δεδομένων και ειδικά το Λογισμικό Εφαρμοσμένης Μηχανικής αποτελεί κύρια πηγή δεδομένων με κώδικα, βιβλιοθήκες συναρτήσεων, αντικείμενα κλπ., των οποίων η διαχείριση και η διατήρηση απαιτεί τη χρησιμοποίηση ισχυρών εργαλείων.

Εικονικοί Κόσμοι (Virtual Worlds): Πολλές εφαρμογές χρησιμοποιούν τρις-διάστατους εικονικούς χώρους με αποτέλεσμα να υπάρχει διαθέσιμη αξιοσημείωτη ποσότητα πηγών πληροφορίας για αντικείμενα και χώρους εικονικής πραγματικότητας. Η διαχείριση τέτοιων πηγών βρίσκεται σε ερευνητικό στάδιο, ωστόσο το μέγεθος των συλλογών τέτοιας πληροφορίας συνεχίζει να αυξάνεται.

Εκθέσεις κειμένου, υπομνήματα και μηνύματα ηλεκτρονικού ταχυδρομείου (Text reports, memos and e-mail messages): Το μεγαλύτερο μέρος της επικοινωνίας εντός και μεταξύ εταιριών ή ερευνητικών οργανισμών ακόμα και των απλών ανθρώπων μεταξύ τους, βασίζεται σε εκθέσεις κειμένου και υπομνήματα τα οποία ανταλλάσσουν μέσω e-mails. Τα μηνύματα αυτά αποθηκεύονται σε ψηφιακή μορφή για μελλοντική χρήση και αναφορά και δημιουργούν ψηφιακές βιβλιοθήκες.

Οι πηγές πληροφορίας του Παγκόσμιου Ιστού (The World Wide Web Repositories): Με το ξεκίνημα του Παγκόσμιου Ιστού το 1993, έγγραφα ποικίλων μορφών, περιεχομένου και περιγραφής έχουν συλλεχθεί και διασυνδεθεί με υπερ-συνδέσμους με αποτέλεσμα να αποτελούν τη μεγαλύτερη πηγή αποθηκευμένης πληροφορίας που έχει κατασκευαστεί έως τώρα. Ο Παγκόσμιος Ιστός παρά τη δυναμική και μη δομημένη φύση του, τα ετερογενή χαρακτηριστικά του, την ασυνέπεια και τον πλεονασμό, αποτελεί την πιο σημαντική συλλογή δεδομένων που χρησιμοποιείται για αναφορά, λόγω της ποικιλίας των θεμάτων που καλύπτει και τις άπειρες συνεισφορές πηγών πληροφορίας και εκδοτών. Πολλοί πιστεύουν ότι ο Παγκόσμιος Ιστός θα αποτελέσει τη συλλογή της ανθρώπινης γνώσης.

Γενικά, τα είδη της αποθηκευμένης πληροφορίας είναι ουσιαστικά ανεξάντλητα και επομένως η εξόρυξη γνώσης θα πρέπει να είναι εφαρμόσιμη σε οποιοδήποτε είδος δεδομένων. Ωστόσο, οι αλγόριθμοι μπορεί να διαφέρουν όταν εφαρμόζονται σε διαφορετικούς τύπους δεδομένων. Γενικά, η εξόρυξη γνώσης χρησιμοποιείται για τύπους δεδομένων που παρουσιάζουν σημαντικές διαφορές και αποθηκεύονται με ποικίλες μορφές όπως, απλά αρχεία, σχεσιακές βάσεις δεδομένων, αποθήκες δεδομένων, μη δομημένες πηγές πληροφορίας (ο Παγκόσμιος Ιστός), βάσεις χωρικών δεδομένων, βάσεις χρονολογικών δεδομένων κλπ. Στη συνέχεια δίνονται κάποια σχετικά παραδείγματα με μεγαλύτερη λεπτομέρεια.

Απλά Αρχεία (Flat files): Τα αρχεία αυτά αποτελούν τις πιο συνηθισμένες πηγές δεδομένων για τους αλγόριθμους εξόρυξης γνώσης, κυρίως στο επίπεδο της έρευνας. Πρόκειται για απλά αρχεία κειμένου ή αρχεία δυαδικής μορφοποίησης και τα δεδομένα που περιέχουν είναι συνήθως στοιχεία συναλλαγών, χρονολογικά δεδομένα, επιστημονικές μετρήσεις κλπ.

Σχεσιακές Βάσεις Δεδομένων (Relational Data Bases): Οι βάσεις αυτές αποτελούνται από σύνολα πινάκων, με γραμμές και στήλες, που περιέχουν τιμές των χαρακτηριστικών των οντοτήτων ή τιμές των χαρακτηριστικών των συσχετίσεων των οντοτήτων. Η γλώσσα που χρησιμοποιείται στις σχεσιακές βάσεις δεδομένων είναι η SQL η οποία επιτρέπει την ανάκτηση και τον έλεγχο των αποθηκευμένων δεδομένων στους πίνακες. Οι αλγόριθμοι εξόρυξης γνώσης προσαρμόζονται με μεγαλύτερη ευκολία στις σχεσιακές βάσεις δεδομένων απ' ότι στα απλά αρχεία, κυρίως λόγω της κανονικοποιημένης δομής τους.

Αποθήκες Δεδομένων (Data Warehouses): Η Αποθήκη Δεδομένων είναι ουσιαστικά μια βάση δεδομένων στην οποία συλλέγονται δεδομένα από πολλές πηγές, συχνά ετερογενείς, με σκοπό την ανάλυση τους ως σύνολο κάτω από το ίδιο ενοποιημένο σχήμα. Τα ετερογενή δομένα φορτώνονται, καθαρίζονται, μετασχηματίζονται και τελικά συγκεντρώνονται σε μια Αποθήκη με πολυδιάστατη δομή στα δεδομένα της οποίας μπορεί να εφαρμοστεί OLAP επεξεργασία. Οι OLAP κύβοι περιέχουν σύνθετα μέλη και διαστάσεις με ιεραρχίες, με αποτέλεσμα, σε αντίθεση με τις σχεσιακές βάσεις, να μη διατηρούν μεγάλη λεπτομέρεια για τα δεδομένα. Αυτό συμβαίνει κατά τη διαδικασία δημιουργίας συναθροίσεων. Επομένως, είναι δύσκολη η ανακάλυψη κρυμμένης γνώσης ανάμεσα σε τέτοια μέλη και διαστάσεις. Το πρόβλημα μπορεί να ξεπεραστεί με τη χρησιμοποίηση μοντέλων εξόρυξης γνώσης πάνω σε OLAP πηγές. Στον Πίνακα 2-2 παρουσιάζεται τι είναι εφικτό και τι όχι με το OLAP και τα μοντέλα εξόρυξης γνώσης (Data Mining Models).

Βάσεις Δεδομένων Συναλλαγών (Transaction Databases): Μια βάση συναλλαγών είναι ένα σύνολο εγγραφών – συναλλαγών, που η κάθε μία συμβαίνει σε κάποια χρονική στιγμή, έχει ένα αναγνωριστικό και ένα σύνολο αντικειμένων. Επειδή, οι σχεσιακές βάσεις δεν επιτρέπουν τη δημιουργία εμφωλευμένων πινάκων (π.χ. ένα σύνολο δομένων να αποτελεί την τιμή ενός χαρακτηριστικού), οι συναλλαγές συνήθως αποθηκεύονται σε απλά αρχεία ή σε δύο κανονικοποιημένους πίνακες συναλλαγών, έναν για τις συναλλαγές και έναν για το σύνολο των αντικειμένων. Η τεχνική εξόρυξης γνώσης που εφαρμόζεται σε τέτοια δεδομένα ονομάζεται ανάλυση δεδομένων από το «καλάθι της νοικοκυράς» (market-basket analysis) ή αλλιώς κανόνες συσχετίσεων (association rules) αφού ουσιαστικά γίνεται προσπάθεια μελέτης των κρυμμένων συσχετίσεων ανάμεσα στα αντικείμενα.

Βάσεις Πολυμέσων (Multimedia Databases): Οι βάσεις αυτές περιέχουν βίντεο, φωτογραφίες και γενικά οπτικοακουστικά δεδομένα. Η εξόρυξη γνώσης από πηγές πολυμέσων απαιτεί τη χρησιμοποίηση τεχνολογιών αναγνώρισης προτύπων, αναγνώρισης φωνής, γραφικά με υπολογιστές κλπ.

Βάσεις Χωρικών Δεδομένων (Spatial Databases): Εδώ αποθηκεύεται γεωγραφική πληροφορία, όπως χάρτες κλπ. Τα χωρικά δεδομένα είναι δεδομένα τα οποία έχουν μια χωρική συνιστώσα ή συνιστώσα θέσης. Είναι, κατά μία έννοια, δεδομένα αντικειμένων που βρίσκονται σε ένα φυσικό χώρο ο οποίος δηλώνεται ρητά με ένα ή και περισσότερα γνωρίσματα θέσης, όπως η διεύθυνση ή το γεωγραφικό πλάτος/μήκος. Κλασσικοί αλγόριθμοι εξόρυξης γνώσης βρίσκουν εφαρμογές και σε τέτοια δεδομένα, ωστόσο έχουν αναπτυχθεί νέες τεχνικές που προσαρμόζονται καλύτερα στις ιδιαιτερότητες των χωρικών δεδομένων.

Βάσεις Χρονολογικών Δεδομένων (Time-Series Databases): Οι βάσεις δεδομένων δεν περιέχουν συνήθως χρονολογικά δεδομένα καθώς τα δεδομένα τους αφορούν σε ένα συγκεκριμένο σημείο στο χρόνο. Οι βάσεις χρονολογικών δεδομένων περιέχουν δεδομένα που σχετίζονται με το χρόνο και η συνεχής ροή νέων δεδομένων προς αυτές δημιουργεί την ανάγκη ανάλυσης σε πραγματικό χρόνο. Η εξόρυξη γνώσης σε τέτοια δεδομένα περιλαμβάνει τη μελέτη των τάσεων και των συσχετίσεων που παρουσιάζονται ενδιάμεσα της εξέλιξης των διαφορετικών μεταβλητών καθώς και την πρόβλεψη των τάσεων και τις κινητικότητας των μεταβλητών σε βάθος χρόνου.

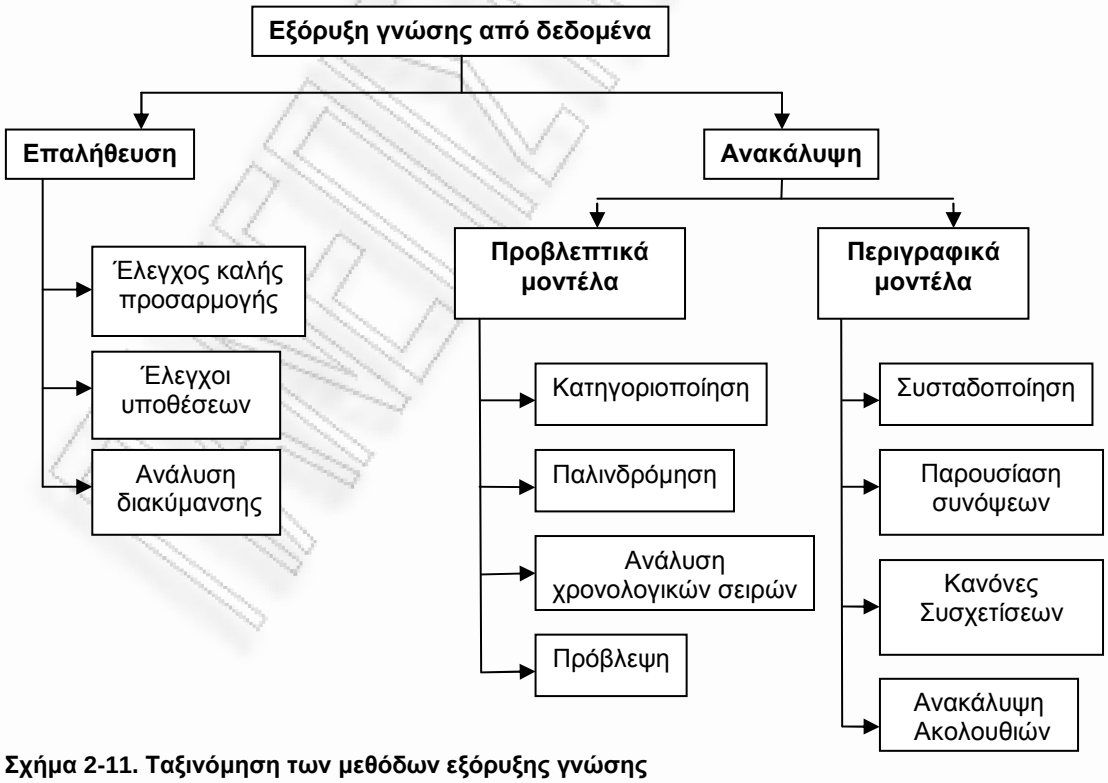
Ο Παγκόσμιος Ιστός (The World Wide Web): Αποτελεί τη μεγαλύτερη διαθέσιμη ετερογενή πηγή πληροφορίας. Ο Παγκόσμιος Ιστός συγκεντρώνει δεδομένα που προέρχονται από αναρίθμητες πηγές πληροφορίας, συγγραφείς και εκδότες, ενώ παράλληλα παρέχει καθημερινά τη δυνατότητα σε εκατομμύρια χρήστες να έχουν πρόσβαση στα δεδομένα αυτά. Τα δεδομένα στον Παγκόσμιο Ιστό είναι οργανωμένα σε διασυνδεδεμένα έγγραφα διαφόρων μορφοποιήσεων, όπως κείμενα, ήχος, εικόνα, βίντεο, πρωτογενή δεδομένα ή ακόμα και εφαρμογές. Ουσιαστικά, τον Παγκόσμιο Ιστό συνθέτουν τρεις συνιστώσες: το περιεχόμενο του Ιστού, δηλαδή τα διαθέσιμα έγγραφα, η δομή του Ιστού που περιλαμβάνει τους υπερ-συνδέσμους και τις σχέσεις μεταξύ των εγγράφων και τέλος η χρήση του Ιστού, δηλαδή πώς και πότε προσπελαύνονται οι διάφορες πηγές. Επομένως, η εξόρυξη γνώσης στον Παγκόσμιο Ιστό διακρίνεται σε εξόρυξη γνώσης από το περιεχόμενο του Παγκόσμιου Ιστού (web content mining), σε εξόρυξη γνώσης από τη δομή του Παγκόσμιου Ιστού (web structure mining) και σε εξόρυξη γνώσης από τη χρήση του Παγκόσμιου Ιστού (web usage mining).

Πίνακας 2-2. Διαφορές OLAP και Data Mining

OLAP	DATA MINING
Εστιάζει σε ιστορικά δεδομένα	Εστιάζει σε μελλοντικά αποτελέσματα ή τάσεις
Συναθροίζει δεδομένα με τη χρήση προκαθορισμένης ομαδοποίησης	Προϋποθέτει λεπτομερή δεδομένα
Επαλήθευση/Αντικειμενικά αποτελέσματα	Ανακάλυψη
Ad-hoc ερωτήματα και εκθέσεις	Τεχνικές Στατιστικής και Μηχανικής Μάθησης
Περιορισμένη δυνατότητα εξαγωγής αξιόπιστων εκτιμήσεων με προβλέψεις	Μοντέλα δεδομένων διαθέσιμα για πρόβλεψη, ανακάλυψη προτύπων, εκτίμηση και παραγωγή σωστών αποτελεσμάτων για ανάλυση τάσεων και διενέργεια προβλέψεων
Το OLAP μπορεί να χρησιμοποιηθεί ως πηγή δεδομένων για τα μοντέλα εξόρυξης γνώσης	Τα αποτελέσματα των μοντέλων εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν σε OLAP εφαρμογές ως νέες μεταβλητές προς πρόβλεψη ή ως χαρακτηριστικά

2.2.4 Γενική Αναφορά στις Μεθόδους Εξόρυξης Γνώσης από Δεδομένα

Ο τεράστιος όγκος δεδομένων που αποθηκεύονται σε ποικίλες μορφές οδήγησε και στην ανάπτυξη πολλών μεθόδων εξόρυξης γνώσης με διαφορετικούς σκοπούς και στόχους. Στο παρακάτω Σχήμα παρουσιάζεται μια γενική ταξινόμηση των μεθόδων εξόρυξης γνώσης από τα δεδομένα.



Σχήμα 2-11. Ταξινόμηση των μεθόδων εξόρυξης γνώσης

Όπως φαίνεται στο παραπάνω Σχήμα, υπάρχουν δύο βασικοί τύποι εξόρυξης γνώσης: η *επαλήθευση* και η *ανακάλυψη*. Κατά την επαλήθευση το σύστημα καλείται να επιβεβαιώσει τις υποθέσεις που έχει κάνει ο χρήστης ενώ κατά την *ανακάλυψη* το σύστημα καλείται να βρει νέους κανόνες και πρότυπα μέσα από αυτόνομες διαδικασίες. Οι μέθοδοι επαλήθευσης ασχολούνται κυρίως με την εκτίμηση μιας υπόθεσης που προτείνεται από μια εξωτερική πηγή. Πρόκειται ουσιαστικά για τις παραδοσιακές μεθόδους της Στατιστικής (π.χ. έλεγχος καλής προσαρμογής, έλεγχος υποθέσεων, ανάλυση διακύμανσης) οι οποίες σχετίζονται λιγότερο με την εξόρυξη γνώσης από δεδομένα συγκριτικά με τις μεθόδους ανακάλυψης. Οι μέθοδοι ανακάλυψης είναι αυτές που εντοπίζουν αυτόματα πρότυπα στα δεδομένα και μας ενδιαφέρουν περισσότερο για την ανάπτυξη της παρούσας εργασίας.

Το επόμενο στάδιο της ανακάλυψης χωρίζεται στη δημιουργία δύο μοντέλων του *προβλεπτικού* και του *περιγραφικού*.

Το προβλεπτικό μοντέλο (predictive model) έχει ως στόχο την πρόγνωση της τιμής μιας μεταβλητής μέσα από τις τιμές άλλων μεταβλητών που είναι γνωστές. Δηλαδή επιδιώκει να κάνει κάποια πρόβλεψη για τις τιμές των δεδομένων με χρησιμοποίηση γνωστών αποτελεσμάτων που έχει βρει από άλλα δεδομένα. Ως εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούνται για τη μοντελοποίηση μιας πρόβλεψης αναφέρονται η κατηγοριοποίηση, η παλινδρόμηση, η ανάλυση χρονολογικών σειρών και η πρόβλεψη.

Το περιγραφικό μοντέλο (descriptive model) έχει ως στόχο την περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα αναγνωρίζοντας πρότυπα ή συσχετίσεις ανάμεσα στα δεδομένα. Το περιγραφικό μοντέλο, σε αντίθεση με το προβλεπτικό, στοχεύει στην ερμηνεία των δεδομένων ερευνώντας τις ιδιότητές τους, τις συσχετίσεις που υπάρχουν ανάμεσά τους, χωρίς να προβλέπει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών είναι περιγραφικές εργασίες εξόρυξης γνώσης από δεδομένα.

Η εξόρυξη γνώσης μπορεί να χρησιμοποιηθεί για την επίλυση εκατοντάδων επιχειρησιακών προβλημάτων. Στη συνέχεια ακολουθεί μια συνοπτική αναφορά στις εργασίες εξόρυξης γνώσης των δύο μοντέλων ανακάλυψης βάσει τους Σχήματος 2-11.

Κατηγοριοποίηση (classification): Περιλαμβάνει την κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο σε μια από ένα σύνολο από προκαθορισμένες κατηγορίες – κλάσεις (classes). Αναφέρεται και ως εποπτευόμενη μάθηση, επειδή οι κατηγορίες καθορίζονται πριν εξεταστούν τα δεδομένα. Στους αλγόριθμους κατηγοριοποίησης οι κατηγορίες πρέπει να ορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων (δεδομένα εκπαίδευσης) και τα νέα δεδομένα να κατηγοριοποιούνται με βάση της γνώση που παρέχουν τα δεδομένα εκπαίδευσης.

Παλινδρόμηση (regression): Ένας από τους κύριους σκοπούς της προσαρμογής καμπυλών είναι η εκτίμηση τη εξαρτημένης μεταβλητής από την ανεξάρτητη μεταβλητή. Η μέθοδος ή η διαδικασία εκτίμησης ονομάζεται παλινδρόμηση. Ουσιαστικά, χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή υπό πρόβλεψη. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συναρτήσεων (π.χ. γραμμική, λογαριθμική κλπ) και καθορίζει τη συνάρτηση που μοντελοποιεί καλύτερα τα δεδομένα που έχουν δοθεί. Η κύρια διαφορά της παλινδρόμησης με την κατηγοριοποίηση είναι ότι το υπό πρόβλεψη χαρακτηριστικό παίρνει συνεχείς τιμές.

Ανάλυση χρονολογικών σειρών ή χρονοσειρών (time series analysis): Μελετά την τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο με κάποια περιοδικότητα (π.χ. ημερήσια, εβδομαδιαία, μηνιαία κλπ.). Ως χρονολογική σειρά ορίζεται η ακολουθία των τιμών μιας μεταβλητής οι οποίες λαμβάνονται σε προκαθορισμένα χρονικά σημεία που συνήθως ισαπέχουν ή αναφέρονται σε διαδοχικές περιόδους ίδιας διάρκειας. Η γραφική απεικόνιση των χρονολογικών σειρών, οι οποίες εκφράζονται σε απόλυτα ή σε σχετικά μεγέθη, γίνεται βάσει ειδικών διαγραμμάτων, τα λεγόμενα χρονογράμματα ή χρονοδιαγράμματα. Υπάρχουν τρεις βασικές λειτουργίες που χρησιμοποιούνται στην ανάλυση χρονοσειρών. Η πρώτη περιλαμβάνει τη χρησιμοποίηση μονάδων μέτρησης απόστασης ώστε να καθοριστούν οι ομοιότητες ανάμεσα σε διαφορετικές χρονοσειρές. Η δεύτερη λειτουργία εξετάζει τη δομή της χρονοσειράς για να κατηγοριοποιήσει τη συμπεριφοράς της. Τέλος, η τρίτη λειτουργία χρησιμοποιεί διαγράμματα χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

Πρόβλεψη (prediction): Αναφέρεται στην πρόβλεψη ελλιπών αριθμητικών τιμών ή στην αύξηση/μείωση τάσεων που σχετίζονται με δεδομένα ως προς τον χρόνο. Η πρόβλεψη μπορεί

να θεωρηθεί ως κατηγοριοποίηση καθώς η κυρίαρχη ιδέα είναι η χρησιμοποίηση μεγάλου αριθμού τιμών δεδομένων προηγούμενων περιόδων για απόδοση πιθανών μελλοντικών τιμών στα δεδομένα. Δηλαδή οι εφαρμογές πρόβλεψης επιδιώκουν την απόδοση μιας τιμής σε μια μελλοντική κατάσταση παρά σε μια τρέχουσα.

Συσταδοποίηση (clustering): Είναι παρόμοια με την κατηγοριοποίηση, δηλαδή επιχειρεί να βρει ομάδες – συστάδες (clusters) παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα γνωρίσματα των χαρακτηριστικών που περιλαμβάνουν. Η κύρια διαφορά της με την κατηγοριοποίηση είναι ότι οι συστάδες δεν είναι προκαθορισμένες αλλά ορίζονται από τα ίδια τα δεδομένα. Αναφέρεται και ως μη εποπτευόμενη μάθηση καθώς η κλάση στην οποία ανήκουν τα δεδομένα εκπαίδευσης δεν είναι εκ των προτέρων γνωστή. Η βασική αρχή που διέπει όλες τις προσεγγίσεις συσταδοποίησης βασίζεται στη μεγιστοποίηση της ομοιότητας μεταξύ αντικειμένων που ανήκουν στην ίδια ομάδα (intra-class similarity) και στην ελαχιστοποίηση της ομοιότητας μεταξύ αντικειμένων διαφορετικών ομάδων (inter-class similarity).

Παρουσίαση συνόψεων (summarization) ή Χαρακτηρισμός (characterization): Απεικονίζει τα δεδομένα σε υποσύνολά τους με συνοδευτικές απλές περιγραφές και χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων. Εξάγει αντιπροσωπευτικές πληροφορίες για τη βάση δεδομένων και παράγει τους ονομαζόμενους *χαρακτηριστικούς κανόνες (characteristic rules)*. Καθώς ένα κύβος δεδομένων περιέχει συγκεντρωμένα δεδομένα, οι απλές λειτουργίες OLAP ταιριάζουν στον σκοπό της παρουσίασης συνόψεων.

Κανόνες Συσχετίσεων (Association Rules): Ένας κανόνας συσχετίσεων είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ διαφορετικών χαρακτηριστικών σε ένα σύνολο δεδομένων. Οι κανόνες συσχετίσεων βρίσκουν εφαρμογή στην ανάλυση του «καλαθιού αγοράς» (market basket analysis) καθώς σκοπός τους είναι η αναγνώριση προϊόντων που συνήθως αγοράζονται μαζί.

Ανακάλυψη ακολουθιών (sequence discovery): Χρησιμοποιείται για τον καθορισμό προτύπων σε σειριακά δεδομένα. Σειρές διακριτών τιμών ή καταστάσεων γνωρισμάτων δεδομένων συνθέτουν μια ακολουθία. Τα δεδομένα τόσο στην ανακάλυψη ακολουθιών όσο και στην ανάλυση χρονολογικών σειρών περιέχουν γειτονικές παρατηρήσεις που αλληλεξαρτώνται, με τη μόνη διαφορά ότι στην πρώτη περίπτωση τα δεδομένα είναι διακριτά ενώ στη δεύτερη είναι συνεχή. Επίσης, η διαφορά της ανακάλυψης ακολουθιών με τους κανόνες συσχετίσεων έγκειται στο γεγονός ότι τα μοντέλα ακολουθίας θεωρούν ότι τα προϊόντα αγοράζονται με κάποια σειρά ενώ τα μοντέλα συσχετίσεων θεωρούν ότι κάθε προϊόν έχει την ίδια πιθανότητα να αγοραστεί και δεν εξαρτάται από τις άλλες αγορές.

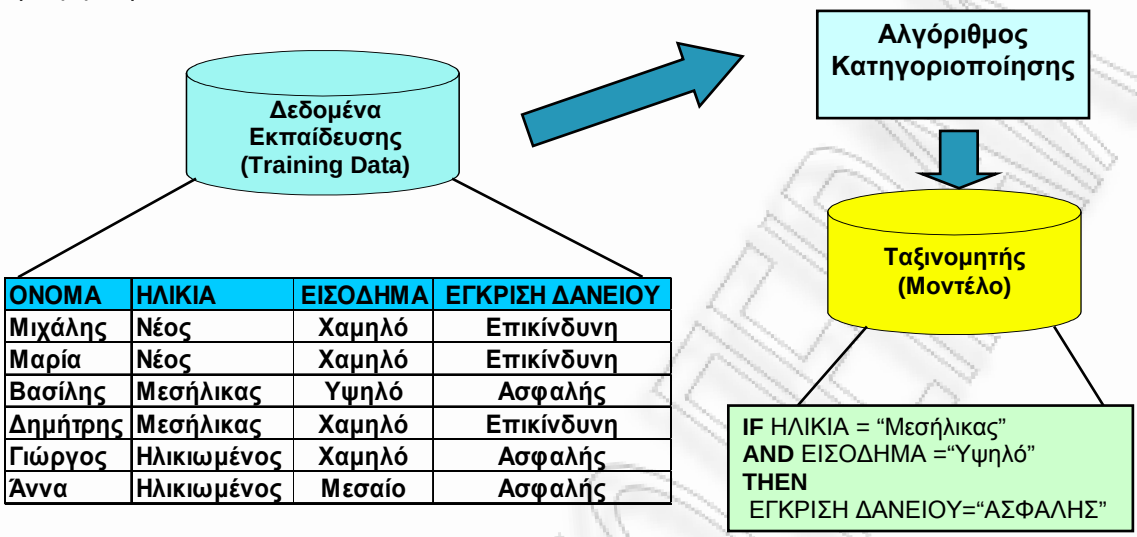
2.2.5 Εργασίες Εξόρυξης Γνώσης από Δεδομένα

Στην προηγούμενη παράγραφο έγινε μια συνοπτική αναφορά στις βασικές μεθόδους εξόρυξης γνώσης από δεδομένα και στις σχετικές εργασίες τους (προβλεπτικές, περιγραφικές). Στη συνέχεια θα παρουσιαστεί μια πιο αναλυτική προσέγγιση του τρόπου λειτουργίας και δομής των εργασιών εξόρυξης γνώσης που θεωρούνται ως πιο σημαντικές, οι οποίες χρησιμοποιούν κανόνες Μηχανικής Μάθησης, και είναι οι εξής: Κατηγοριοποίηση, Συσταδοποίηση, Κανόνες Συσχετίσεων.

Κατηγοριοποίηση (classification): Περιλαμβάνει την εκμάθηση μιας τεχνικής, δηλαδή την κατασκευή ενός μοντέλου, που να προβλέπει την κατηγορία – κλάση ενός στοιχείου από προκαθορισμένες τιμές. Πρόκειται για εποπτευόμενη μάθηση, καθώς αρχικά δίνεται ένα σύνολο από εγγραφές (σύνολο εκπαίδευσης – training set), εκ των οποίων η κάθε μια συνοδεύεται από μια ετικέτα που δείχνει την κλάση στην οποία ανήκει. Στη συνέχεια, οι νέες εγγραφές κατηγοριοποιούνται με βάση τη γνώση που παρέχει το σύνολο εκπαίδευσης.

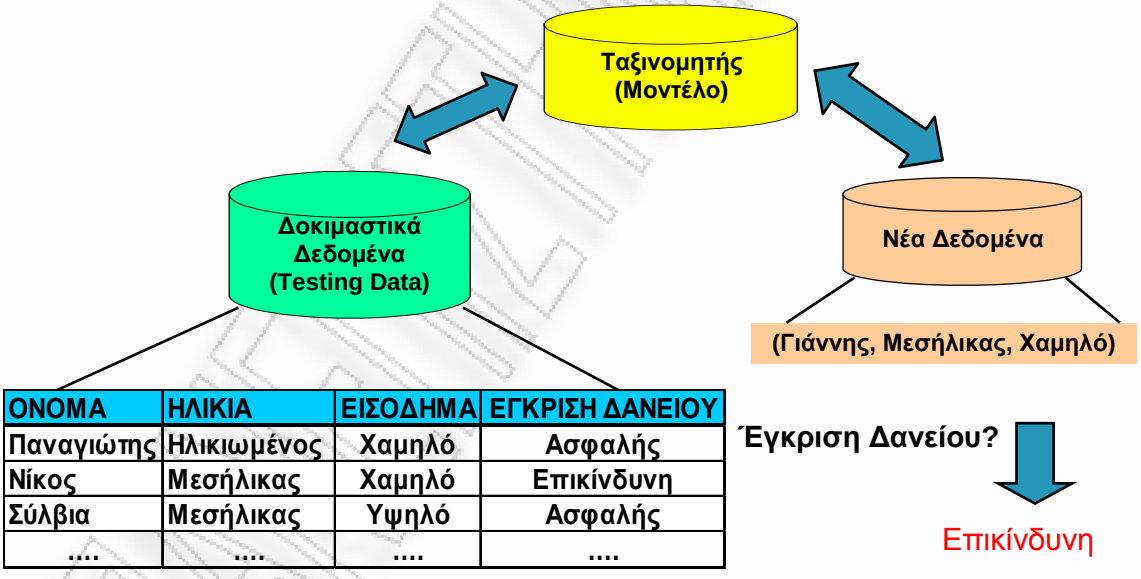
Οι αλγόριθμοι κατηγοριοποίησης εφαρμόζονται σε διακριτά δεδομένα τα οποία έχουν προταξινομηθεί σε συγκεκριμένες κλάσεις με στόχο την εξαγωγή κανόνων οι οποίοι μπορεί μετέπειτα να χρησιμοποιηθούν για κατηγοριοποίηση νέων δεδομένων στις ίδιες κλάσεις. Κάθε αλγόριθμος κατηγοριοποίησης αναλύεται σε τρεις λειτουργίες: (α) δίνεται ένα σύνολο από δεδομένα σαν είσοδος, (β) ο αλγόριθμος με τη σειρά του «μαθαίνει» από το πώς αυτά τα δεδομένα έχουν κατηγοριοποιηθεί, δηλαδή κατανοεί τους κανόνες βάσει των οποίων κατηγοριοποιήθηκαν τα δεδομένα, και (γ) βάσει των συγκεκριμένων κανόνων έχει την ικανότητα

να κατηγοριοποιήσει τα νέα δεδομένα. Επίσης, ο αλγόριθμος εξάγει ένα σύνολο κανόνων που ονομάζεται ταξινομητής (classifier). Τα Σχήματα 2-12, 2-13 παρουσιάζουν ένα παράδειγμα δημιουργίας μοντέλου κατηγοριοποίησης και εφαρμογής του πάνω σε δεδομένα που αφορούν την έγκριση δανείου.



ΟΝΟΜΑ	ΗΛΙΚΙΑ	ΕΙΣΟΔΗΜΑ	ΕΓΚΡΙΣΗ ΔΑΝΕΙΟΥ
Μιχάλης	Νέος	Χαμηλό	Επικίνδυνη
Μαρία	Νέος	Χαμηλό	Επικίνδυνη
Βασίλης	Μεσήλικας	Υψηλό	Ασφαλής
Δημήτρης	Μεσήλικας	Χαμηλό	Επικίνδυνη
Γιώργος	Ηλικιωμένος	Χαμηλό	Ασφαλής
Άννα	Ηλικιωμένος	Μεσαίο	Ασφαλής

Σχήμα 2-12. Δημιουργία μοντέλου



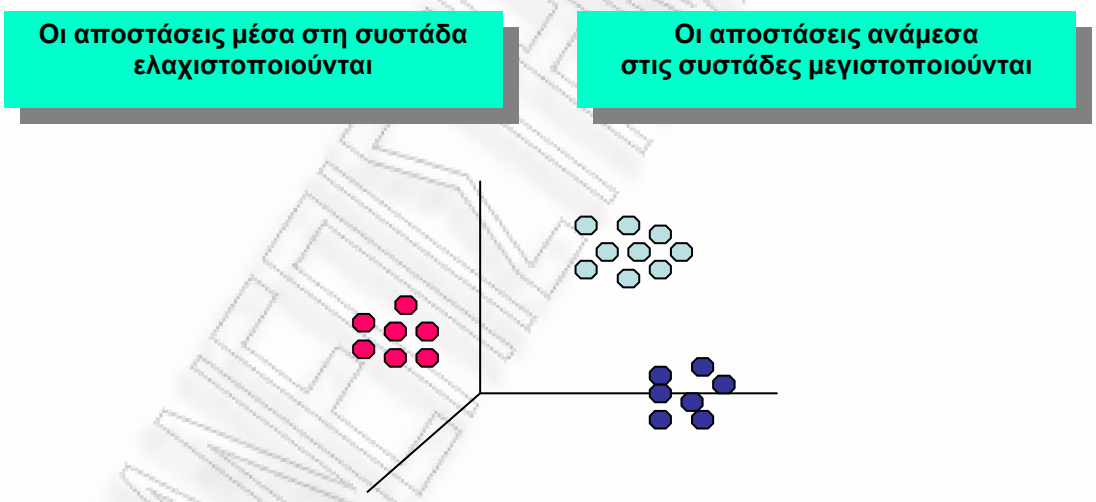
ΟΝΟΜΑ	ΗΛΙΚΙΑ	ΕΙΣΟΔΗΜΑ	ΕΓΚΡΙΣΗ ΔΑΝΕΙΟΥ
Παναγιώτης	Ηλικιωμένος	Χαμηλό	Ασφαλής
Νίκος	Μεσήλικας	Χαμηλό	Επικίνδυνη
Σύλβια	Μεσήλικας	Υψηλό	Ασφαλής
....	....	....	....

Σχήμα 2-13. Εφαρμογή μοντέλου

Η κατηγοριοποίηση ίσως είναι η πιο δημοφιλής και αποτελεσματική τεχνική εξόρυξης γνώσης και βρίσκει εφαρμογές στην αναγνώριση προτύπων και εικόνας, σε ιατρικές διαγνώσεις, στην ανίχνευση λαθών σε βιομηχανικές εφαρμογές καθώς στην ταξινόμηση των τάσεων της οικονομίας. Οι αλγόριθμοι κατηγοριοποίησης διακρίνονται ανάλογα με το είδος του ταξινομητή που παράγουν σε λίστες αποφάσεων, που έχουν τη μορφή λογικών κανόνων, και σε δένδρα αποφάσεων. Οι αλγόριθμοι που χρησιμοποιούνται για κατηγοριοποίηση στον SQL Server 2005, πάνω στον οποίο υλοποιήθηκε η παρούσα εργασία, είναι οι: Decision Trees (Δένδρα Αποφάσεων), Naïve Bayes και Neural Networks (Νευρωνικά Δίκτυα).

Συσταδοποίηση (clustering): Η τεχνική ομαδοποίησης χωρίζει ουσιαστικά ένα σύνολο εγγραφών σε ομάδες – συστάδες έτσι ώστε οι εγγραφές που βρίσκονται στην ίδια συστάδα να έχουν περισσότερες ομοιότητες μεταξύ τους, με βάση ορισμένα προκαθορισμένα κριτήρια, σε σύγκριση με τις εγγραφές των άλλων συστάδων. Πρόκειται για μη εποπτευόμενη μάθηση καθώς αρχικά δίνεται ένα σύνολο από σημεία που το καθένα έχει κάποια γνωρίσματα και μια μέτρηση ομοιότητας μεταξύ τους. Στη συνέχεια τα σημεία αυτά ομαδοποιούνται σε συστάδες που είναι εκ των προτέρων γνωστές, με τρόπο ώστε τα σημεία μιας συστάδας να είναι πιο όμοια μεταξύ τους και τα σημεία διαφορετικών συστάδων να είναι λιγότερο όμοια μεταξύ τους.

Οι αλγόριθμοι συσταδοποίησης χρησιμοποιούν επαναληπτικές μεθόδους για την ομαδοποίηση των εγγραφών ενός συνόλου δεδομένων σε συστάδες με παρόμοια χαρακτηριστικά. Ουσιαστικά, οι αλγόριθμοι πρώτα αναγνωρίζουν σχέσεις μεταξύ των δεδομένων ενός συνόλου και μετά παράγουν μια ακολουθία συστάδων βάσει των σχέσεων αυτών. Μετά την πρώτη εύρεση συστάδων, ο αλγόριθμος συσταδοποίησης υπολογίζει πόσο καλά οι συστάδες παρουσιάζουν την ομαδοποίηση των σημείων και με επανάληψη της διαδικασίας δημιουργεί συστάδες που παρουσιάζουν καλύτερα τα δεδομένα. Η όλη διαδικασία σταματά όταν τα αποτελέσματα δεν μπορούν να βελτιωθούν περισσότερο. Για την επιλογή του κατάλληλου αλγόριθμου απαραίτητη προϋπόθεση είναι η μελέτη των δεδομένων που θα χρησιμοποιηθούν για τον προσδιορισμό κυρίως του κριτηρίου ομοιότητας των εγγραφών μίας ομάδας. Γενικά η τεχνική της συσταδοποίησης διακρίνεται σε *Στατιστική ή Αριθμητική (statistical/numerical clustering)* και σε *Εννοιολογική (conceptual clustering)*. Στην πρώτη περίπτωση χρησιμοποιούνται διάφορα αριθμητικά κριτήρια ομοιότητας. Έτσι οι ομάδες που προκύπτουν περιγράφονται από αριθμητικές τιμές. Στη δεύτερη περίπτωση ο προσδιορισμός των ομάδων βασίζεται στο νόημα και στις έννοιες που τα διάφορα στοιχεία αντιπροσωπεύουν, καθώς περιέχουν κατηγορικές και όχι αριθμητικές τιμές. Για την οπτική παρουσίαση τις ομαδοποίησης των δεδομένων χρησιμοποιούνται διαγράμματα διασποράς (scatter plots) όπως αυτό του Σχήματος 2-14, το οποίο είναι τρις-διάστατο.



Σχήμα 2-14. Παράδειγμα διαγράμματος διασποράς: Το διάγραμμα παρουσιάζει όλες τις περιπτώσεις του συνόλου δεδομένων, καθώς κάθε σημείο απεικονίζει μια περίπτωση. Οι συστάδες ομαδοποιούν τα σημεία του διαγράμματος και παρουσιάζουν τις σχέσεις που ο αλγόριθμος αναγνωρίζει.

Η συσταδοποίηση χρησιμοποιείται σε πολλά πεδία εφαρμογών, συμπεριλαμβανομένων της βιολογίας, ιατρικής, ανθρωπολογίας, μάρκετινγκ και της οικονομίας. Οι αλγόριθμοι που χρησιμοποιούνται για συσταδοποίηση στον SQL Server 2005, είναι οι: Clustering (Συσταδοποίηση) και Sequence Clustering (Ακολουθιακή Συσταδοποίηση).

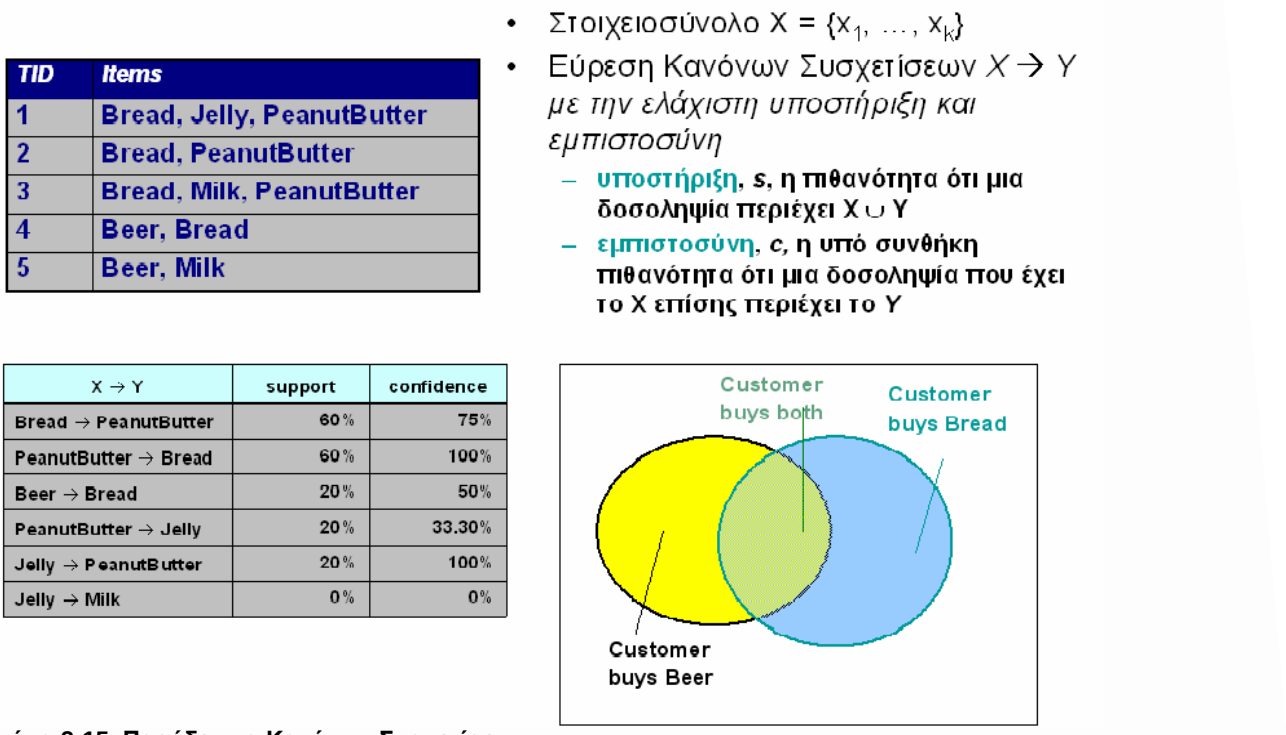
Κανόνες Συσχετίσεων (Association Rules): Η τεχνική αυτή χρησιμοποιείται για την ανακάλυψη προτύπων που περιγράφουν σημαντικές αλληλεξαρτήσεις μεταξύ των διαφόρων πεδίων – χαρακτηριστικών ενός συνόλου δεδομένων. Η πιο συνηθισμένη εφαρμογή των Κανόνων

Συσχετίσεων είναι «η ανάλυση του καλαθιού της νοικοκυράς» καθώς σκοπός τους είναι να αναγνωρισθούν τα αγαθά που αγοράζονται μαζί.

Οι Κανόνες Συσχετίσεων μελετούν το πρόβλημα της εύρεσης συχνών συνόλων αντικειμένων ή στοιχειοσυνόλων (frequent itemsets) σε βάσεις δεδομένων και βασίζονται σε ένα κατώφλι που *ονομάζεται υποστήριξη (support)* το οποίο αναγνωρίζει τα στοιχειοσύνολα. Ένα άλλο κατώφλι είναι η *εμπιστοσύνη (confidence)*, η οποία εκφράζει την υπό συνθήκη πιθανότητα ότι ένα αντικείμενο εμφανίζεται σε μια δοσοληψία όταν επίσης εμφανίζεται ένα άλλο αντικείμενο και χρησιμοποιείται για τον εντοπισμό των κανόνων συσχετίσεων.

Γενικά, ένας κανόνας συσχέτισης ορίζεται πιο αυστηρά ως μία έκφραση της μορφής $X \Rightarrow Y[s, c]$, όπου X, Y είναι σύνολα τιμών των πεδίων, όπως για παράδειγμα σύνολα οικονομικών αγαθών και τα s, c αντιστοιχούν στα κατώφλια υποστήριξη και εμπιστοσύνη. Η υποστήριξη αντιπροσωπεύει τα ποσοστό των δοσοληψιών που περιέχουν το $X \cup Y$ ή διαφορετικά η πιθανότητα $P(X \cup Y)$. Η εμπιστοσύνη είναι η αναλογία του πλήθους των δοσοληψιών που περιέχουν το $X \cup Y$ ως προς το πλήθος των δοσοληψιών που περιέχουν το X ή αλλιώς η πιθανότητα υπό συνθήκη $P(X \cup Y / X) = P(X \cup Y) / P(X)$. Το Σχήμα 2-15 παρουσιάζει ένα παράδειγμα με δοσοληψίες κάποιων προϊόντων και τους εξαγόμενους κανόνες συσχετίσεων με την ελάχιστη υποστήριξη και εμπιστοσύνη.

Το πρόβλημα της εύρεσης όλων των κανόνων συσχετίσεων που πληρούν τις επιθυμητές τιμές υποστήριξης και εμπιστοσύνης μπορεί να διαιρεθεί σε δύο υποπροβλήματα: Πρώτον, στην εύρεση όλων των συνδυασμών των προϊόντων που έχουν υποστήριξη πάνω από την ελάχιστη υποστήριξη. Αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (large itemsets) και όλοι οι υπόλοιποι συνδυασμοί μικρές λίστες από προϊόντα (small itemsets). Δεύτερον, στη χρήση όλων των μεγάλων λιστών από προϊόντα για εξόρυξη των κανόνων συσχετίσεων που ικανοποιούν την ελάχιστη εμπιστοσύνη. Οι κανόνες συσχετίσεων χρησιμοποιούνται από καταστήματα λιανικής πώλησης και βοηθούν στο μάρκετινγκ, στη διαφήμιση, στον έλεγχο του καταλόγου απογραφής κλπ. Οι αλγόριθμοι που χρησιμοποιούνται για εύρεση κανόνων συσχετίσεων στον SQL Server 2005, είναι οι: Association Rules (Κανόνες Συσχετίσεων) και Decision Trees (Δένδρα Αποφάσεων) για μικρούς καταλόγους.



TID	Items
1	Bread, Jelly, PeanutButter
2	Bread, PeanutButter
3	Bread, Milk, PeanutButter
4	Beer, Bread
5	Beer, Milk

- Στοιχειοσύνολο X = {x_1, …, x_k}
- Εύρεση Κανόνων Συσχετίσεων $X \rightarrow Y$ *με την ελάχιστη υποστήριξη και εμπιστοσύνη*
  - **υποστήριξη, s, η πιθανότητα ότι μια δοσοληψία περιέχει X ∪ Y**
  - **εμπιστοσύνη, c, η υπό συνθήκη πιθανότητα ότι μια δοσοληψία που έχει το X επίσης περιέχει το Y**

X → Y	support	confidence
Bread → PeanutButter	60%	75%
PeanutButter → Bread	60%	100%
Beer → Bread	20%	50%
PeanutButter → Jelly	20%	33.30%
Jelly → PeanutButter	20%	100%
Jelly → Milk	0%	0%

Σχήμα 2-15. Παράδειγμα Κανόνων Συσχετίσεων

Συνοψίζοντας, οι τρεις τεχνικές εξόρυξης γνώσης από δεδομένα βρίσκουν εφαρμογές σε ποικίλους τομείς και στοχεύουν στην εύρεση χρήσιμων και δυνητικά χρησιμοποιήσιμων πληροφοριών και προτύπων από τα δεδομένα. Στον παρακάτω πίνακα παρουσιάζονται πιθανά σενάρια εξόρυξης γνώσης και οι αντίστοιχοι αλγόριθμοι του SQL Server 2005 που μπορούν να εφαρμοστούν.

Πίνακας 2-3. Ενδεικτικές εφαρμογές αλγορίθμων του SQL Server 2005

Σενάριο	Αλγόριθμοι
Πρόβλεψη ενός χαρακτηριστικού με διακριτές τιμές, π.χ. πρόβλεψη για το αν ο παραλήπτης στοχευόμενης διαφημιστικής καμπάνιας μέσω ταχυδρομείου θα αγοράσει το προτεινόμενο προϊόν.	Decision Trees, Naïve Bayes, Clustering και Neural Network
Πρόβλεψη ενός χαρακτηριστικού με συνεχείς τιμές, π.χ. πρόβλεψη των πωλήσεων του επόμενου έτους.	Decision Trees
Εύρεση ομάδων κοινών αντικειμένων σε δοσοληψίες, π.χ. χρησιμοποίηση «του καλαθιού της νοικοκυράς» ώστε να προταθούν επιπλέον προϊόντα στον πελάτη.	Association Rules, Decision Trees
Εύρεση ομάδων παρόμοιων αντικειμένων, π.χ. κατάτμηση δημογραφικών δεδομένων σε ομάδες με σκοπό την καλύτερη κατανόηση των σχέσεων μεταξύ των χαρακτηριστικών.	Clustering και Sequence Clustering

3. Αποθήκη Μεταναστευτικών Δεδομένων

Ο σκοπός της παρούσας ενότητας είναι η περιγραφή των διαδικασιών και των βημάτων που ακολουθήθηκαν ώστε να υλοποιηθεί μια Αποθήκη Δεδομένων με στοιχεία για τους ενεργούς αλλοδαπούς ασφαλισμένους των δύο μεγαλύτερων ασφαλιστικών οργανισμών της χώρας, του ΙΚΑ-ΕΤΑΜ και του ΟΑΕΕ/ΤΕΒΕ. Στη συνέχεια δίνεται η αρχιτεκτονική του συστήματος, η οποία βασίστηκε στη θεωρία που αναπτύχθηκε στο Κεφάλαιο 2, ακολουθεί η περιγραφή των πρωτογενών πηγών προέλευσης των στοιχείων και η ενότητα ολοκληρώνεται με τη σχεδίαση της Αποθήκης Δεδομένων πάνω στην οποία θα γίνει η OLAP ανάλυση και θα εφαρμοστούν τεχνικές εξόρυξης γνώσης, θέματα που θα παρουσιαστούν στα κεφάλαια 4, 5.

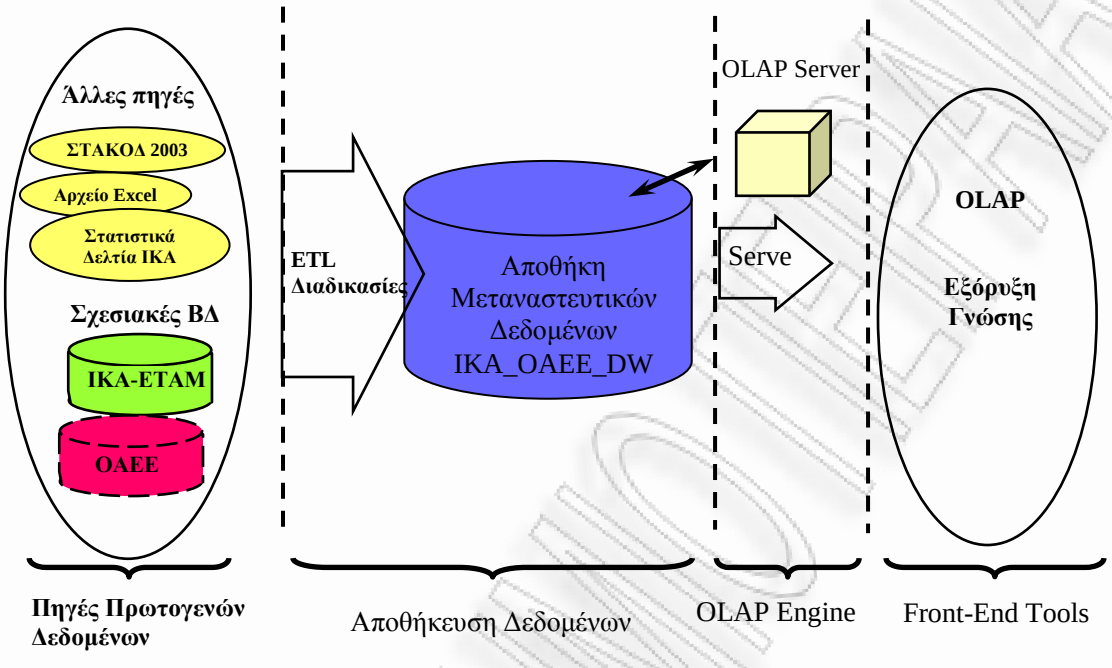
3.1 Αρχιτεκτονική Συστήματος

Σύμφωνα με τον W.H.Inmon [10] , "Η Αποθήκη Δεδομένων (Data Warehouse) αποτελεί ένα θεματο-κεντρικό (subject - oriented), συγκεντρωμένο (integrated), με χρονική διάσταση (time - variant), μη ευμετάβλητο (non – volatile) σύστημα διαχείρισης πληροφοριακών δεδομένων για την υποστήριξη των διαδικασιών λήψης αποφάσεων". Επομένως, στα πλαίσια υλοποίησης μιας Αποθήκης Μεταναστευτικών Δεδομένων η τέσσερις λέξεις κλειδιά, θεματο-κεντρικό, συγκεντρωμένο, με χρονική διάσταση και μη ευμετάβλητο, θα μπορούσαν να αποδοθούν ως εξής:

- Θεματο-κεντρικό: Η συγκεκριμένη Αποθήκη Δεδομένων οργανώνεται γύρω από τους ενεργούς αλλοδαπούς ασφαλισμένους των δύο μεγαλύτερων ασφαλιστικών οργανισμών της χώρας, του ΙΚΑ-ΕΤΑΜ και του ΟΑΕΕ/ΤΕΒΕ, με σκοπό τη μοντελοποίηση και την ανάλυση των διαθέσιμων δημογραφικών και οικονομικών τους δεδομένων για την εξαγωγή συμπερασμάτων σχετικά με τη δραστηριοποίησή τους στον ελλαδικό χώρο.
- Συγκεντρωμένο: Στην Αποθήκη Δεδομένων συγκεντρώθηκαν στοιχεία από πολλαπλές ετερογενείς πηγές. Τα δεδομένα του ΙΚΑ προήλθαν κυρίως από ένα αρχείο EXCEL, που παραχώρησε για τις ανάγκες της εργασίας η Διεύθυνση Στατιστικής του ΙΚΑ_ΕΤΑΜ, ενώ παράλληλα με βάση τα δημοσιευμένα εξαμηνιαία στατιστικά δελτία του ΙΚΑ (Παράρτημα Β.3) παράχθηκαν επιπλέον δεδομένα με διαδικασίες που περιγράφονται στην παράγραφο 3.2.2. Στη συνέχεια, όλη η συγκεντρωμένη πληροφορία ενσωματώθηκε στη σχεσιακή βάση δεδομένων του ΙΚΑ, τα δεδομένα της οποίας φορτώθηκαν με κατάλληλες ETL διαδικασίες στην Αποθήκη Δεδομένων. Στο σημείο αυτό, θα πρέπει να σημειωθεί ότι τόσο η υλοποίηση της σχεσιακής βάσης δεδομένων του ΟΑΕΕ/ΤΕΒΕ όσο και οι ETL διαδικασίες που ακολουθήθηκαν για τη φόρτωση των δεδομένων στην Αποθήκη αποτελούν αντικείμενο υλοποίησης άλλης διπλωματικής εργασίας. Το Σχήμα 3.1 δίνει μια εικόνα των βασικών δομικών στοιχείων της Αποθήκης, της διασύνδεσης των στοιχείων τους και της ροής των δεδομένων.
- Με χρονική διάσταση: Τα δεδομένα που αποθηκεύτηκαν στην Αποθήκη παρέχουν ιστορική πληροφορία για τους αλλοδαπούς ασφαλισμένους του ΙΚΑ κατά τη χρονική περίοδο Ιούνιος 2004 έως Ιούνιος 2008, ενώ για αλλοδαπούς ασφαλισμένους του ΟΑΕΕ/ΤΕΒΕ μόνο για τα έτη 2007 έως 2009.
- Μη ευμετάβλητο: Στην Αποθήκη Δεδομένων εφαρμόστηκε μόνο η λειτουργία της «φόρτωσης» και δεν σημειώθηκε καμιά τροποποίηση στα δεδομένα της, όπως εξ' ορισμού συμβαίνει στα συστήματα Αποθηκών Δεδομένων.

Γενικά, στην Αποθήκη Δεδομένων συγκεντρώθηκαν δημογραφικά και οικονομικά στοιχεία των αλλοδαπών ασφαλισμένων ώστε να είναι η εφικτή η ανάλυση της συνολικής εικόνας του μετανάστη που ζει και δραστηριοποιείται στον ελλαδικό χώρο. Η αδυναμία πλήρους πρόσβασης στα δεδομένα των ασφαλιστικών οργανισμών, καθώς ορισμένα από αυτά αποτελούν ευαίσθητα προσωπικά δεδομένα, καθώς και η τελείως διαφορετική διάθρωση και οργάνωση των πληροφοριακών συστημάτων των δύο οργανισμών, δημιούργησαν μεγάλες δυσκολίες στη συγκέντρωση των δεδομένων σε μια Αποθήκη, οι οποίες ξεπεράστηκαν στο βαθμό που ήταν εφικτό με πολλές παραδοχές κατά τη σχεδίαση και την υλοποίηση του συστήματος. Ωστόσο, το όλο εγχείρημα αποτέλεσε μια ικανοποιητική προσέγγιση του προβλήματος, δηλαδή της ταυτοποίησης και του συγκερασμού ετερογενών δεδομένων σε μια βάση, την αναλυτική

επεξεργασία τους και την εξαγωγή χρήσιμης, άμεσα αξιοποιήσιμης πληροφορίας από τα δεδομένα.



Σχήμα 3-1. Αρχιτεκτονική της Αποθήκης Μεταναστευτικών Δεδομένων

3.2 Σχεδιασμός και Υλοποίηση της Βάσης του ΙΚΑ-ΕΤΑΜ

3.2.1 Πρωτογενείς Πηγές – Ανάλυση Απαιτήσεων - Παραδοχές

Ο σκοπός της παρούσας ενότητας είναι να καθορίσει τις παραδοχές και συμβάσεις που ακολουθήθηκαν στη μοντελοποίηση της βάσης δεδομένων του ΙΚΑ_ΕΤΑΜ λόγω της μη διαθεσιμότητας επαρκών πρωτογενών δεδομένων.

Το σύνολο των μεταναστευτικών δεδομένων (δημογραφικά και οικονομικά δεδομένα) που χρησιμοποιήθηκαν στη βάση προήλθε κυρίως από ένα αρχείο Excel, που παραχωρήθηκε ως αντιπροσωπευτικό δείγμα των αλλοδαπών ασφαλισμένων του ΙΚΑ κατά το δεύτερο εξάμηνο του 2007 (περίπου 65500 εγγραφές) και περιείχε δεδομένα όπως Αύξων Αριθμός, Φύλο, Ημερομηνία Γέννησης, Εθνικότητα, Οικονομική Δραστηριότητα και Νομός κατοικίας-εργασίας. Τα δεδομένα αυτά διατηρήθηκαν αυτούσια στη βάση δεδομένων.

Επίσης, με βάση το έντυπο απογραφής του άμεσα ασφαλισμένου (Παράρτημα Β.1), όπως αυτό διατίθεται από τα τμήματα Μητρώου των υποκαταστημάτων του ΙΚΑ καταγράφηκαν κάποιες απαιτήσεις που περιγράφουν ουσιαστικά το μητρώο ασφαλισμένων του ΙΚΑ, οι οποίες αποτυπώθηκαν στους αντίστοιχους πίνακες της βάσης, ωστόσο όμως δεν αποτελούν ένα ολοκληρωμένο σύνολο λόγω της ανυπαρξίας πρωτογενών δεδομένων. Οι σχετικές απαιτήσεις συνοψίζονται στα παρακάτω:

- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να έχει έναν και μόνον έναν ΑΡΙΘΜΟ ΜΗΤΡΩΟΥ ΑΣΦΑΛΙΣΜΕΝΟΥ - ΑΜΑ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να έχει έναν και μόνον έναν ΤΥΠΟ ΤΑΥΤΟΤΗΤΑΣ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ μπορεί να παρακολουθείται από μία και μόνο μία ΔΟΥ και να έχει έναν και μόνο έναν ΑΡΙΘΜΟ ΦΟΡΟΛΟΓΙΚΟΥ ΜΗΤΡΩΟΥ - ΑΦΜ.

- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να απογραφεί σε ένα και μόνο ένα ΥΠΟΚΑΤΑΣΤΗΜΑ ΤΟΥ ΙΚΑ.
- Ένα ΥΠΟΚΑΤΑΣΤΗΜΑ ΤΟΥ ΙΚΑ μπορεί να απογράφει έναν ή και περισσότερους ΑΣΦΑΛΙΣΜΕΝΟΥΣ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ διαμένει και εργάζεται σε ένα ΝΟΜΟ.
- Σε ένα ΝΟΜΟ μπορεί να διαμένουν και να εργάζονται ένας ή και περισσότεροι ΑΣΦΑΛΙΣΜΕΝΟΙ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να έχει έναν και μόνον έναν ΤΑΧΥΔΡΟΜΙΚΟ ΚΩΔΙΚΑ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να έχει γεννηθεί σε μία και μόνο μία ΧΩΡΑ.
- Σε μία ΧΩΡΑ μπορεί να έχουν γεννηθεί ένας ή και περισσότεροι ΑΣΦΑΛΙΣΜΕΝΟΙ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ πρέπει να παίρνει την εθνικότητα μίας και μόνον μίας ΧΩΡΑΣ.
- Μία ΧΩΡΑ μπορεί να δίνει εθνικότητα σε έναν ή και περισσότερους ΑΣΦΑΛΙΣΜΕΝΟΥΣ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ μπορεί να ήταν ασφαλισμένος σε έναν ή και περισσότερους ΦΟΡΕΙΣ ΤΗΣ ΑΛΛΟΔΑΠΗΣ.
- Ένας ΦΟΡΕΑΣ ΤΗΣ ΑΛΛΟΔΑΠΗΣ μπορεί να παρείχε ασφάλιση σε έναν ή περισσότερους ΑΣΦΑΛΙΣΜΕΝΟΥΣ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ μπορεί να έχει ένα ή και περισσότερα ΙΣΤΟΡΙΚΑ ΑΣΦΑΛΙΣΜΕΝΟΥ.
- Ένα ΙΣΤΟΡΙΚΟ ΑΣΦΑΛΙΣΜΕΝΟΥ πρέπει να υπάρχει για έναν και μόνον έναν ΑΣΦΑΛΙΣΜΕΝΟ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ μπορεί να ήταν ασφαλισμένος σε έναν ή και περισσότερους ΦΟΡΕΙΣ ΤΗΣ ΕΛΛΑΔΑΣ.
- Ένας ΦΟΡΕΑΣ ΤΗΣ ΕΛΛΑΔΑΣ μπορεί να παρείχε ασφάλιση σε έναν ή και περισσότερους ΑΣΦΑΛΙΣΜΕΝΟΥΣ.
- Ένας ΑΣΦΑΛΙΣΜΕΝΟΣ μπορεί να έχει σχέση με έναν ή και περισσότερους ΑΣΦΑΛΙΣΜΕΝΟΥΣ, δηλαδή να έχει ΠΡΟΣΤΑΤΕΥΟΜΕΝΑ ΜΕΛΗ.

Παράλληλα, παράχθηκε πληροφορία σχετικά με την οικονομική δραστηριότητα του εξεταζόμενου αντιπροσωπευτικού συνόλου των αλλοδαπών εργαζομένων του ΙΚΑ για μια περίοδο ετών, βάσει επίσημων δημοσιευμένων στατικών δελτίων του ΙΚΑ, καθώς και του έτους απογραφής τους στον συγκεκριμένο ασφαλιστικό φορέα. Η παραγωγή των δεδομένων οικονομικής δραστηριότητας και έτους απογραφής, δηλαδή του έτους έναρξης ασφάλισης στο ΙΚΑ, περιγράφονται αναλυτικά στην ενότητα 3.2.2. Για κάθε ασφαλισμένο θεωρήθηκε ότι απασχολείται σε μία και μόνο μία οικονομική δραστηριότητα ανά εξάμηνο και έτος, ενώ στην πραγματικότητα κάθε ασφαλισμένος μπορεί να απασχολείται σε περισσότερους του ενός εργοδότες και να εξασκεί διαφορετικές δραστηριότητες. Επίσης, για την απλοποίηση της όλης διαδικασίας θεωρήθηκε ότι ο τόπος κατοικίας και εργασίας ταυτίζονται με το νομό, καθώς δεν υπήρχαν αντίστοιχα διαθέσιμα αναλυτικά στοιχεία.

Τέλος, δημιουργήθηκε υποδομή (αντίστοιχα πεδία στους πίνακες) στη βάση δεδομένων ώστε να υποδεχθεί δεδομένα όταν αυτά θα είναι διαθέσιμα. Στην ενότητα 3.2.3, δίνεται το σχεσιακό σχήμα της βάσης δεδομένων και θα σχολιαστούν αναλυτικά τόσο τα πεδία των πινάκων που περιέχουν δεδομένα όσο και αυτά που δημιουργήθηκαν για μελλοντική χρήση. Ωστόσο, σε κάθε περίπτωση τα δεδομένα που εισήχθησαν στη βάση δεν μπορούν να θεωρηθούν ως ευαίσθητα καθώς κανείς αλλοδαπός δεν μπορεί να ταυτοποιηθεί και κατά συνέπεια να αναγνωριστούν τα στατιστικά δεδομένα που τον αφορούν.

3.2.2 Παραγωγή Δεδομένων

Η παραγωγή των δεδομένων οικονομικής δραστηριότητας και έτους απογραφής έγινε για το αρχικό αντιπροσωπευτικό δείγμα του αρχείου Excel. Ουσιαστικά, υλοποιήθηκαν δύο προγράμματα στο Matlab για την παραγωγή στοιχείων οικονομικής δραστηριότητας για εννέα εξάμηνα (Ιούνιος 2004 έως Ιούνιος 2008) και ένα πρόγραμμα, επίσης στο Matlab, για την παραγωγή έτους απογραφής για τους ασφαλισμένους του αρχείου test_december_07.xls.

Το πρώτο πρόγραμμα που υλοποιήθηκε είναι το change.m (Παράρτημα Γ.1). Στο πρόγραμμα αυτό πραγματοποιήθηκε η ομαδοποίηση της οικονομικής δραστηριότητας σε 17

κατηγορίες σύμφωνα με το ΣΤΑΚΟΔ 2003 (Παράρτημα Β.2 - Στατιστική Ταξινόμηση των Κλάδων της Οικονομικής Δραστηριότητας όπως δίνεται από την Εθνική Στατιστική Υπηρεσία και χρησιμοποιείται στην κωδικοποίηση του ΙΚΑ). Τα πρωτογενή δεδομένα του αρχείου Excel περιείχαν έως και 99 κωδικούς οικονομικής δραστηριότητας ενώ στους εξαμηνιαίους στατιστικούς πίνακες του ΙΚΑ η οικονομική δραστηριότητα απεικονίζεται με τις 17 γενικές κατηγορίες του ΣΤΑΚΟΔ 2003. Έτσι λοιπόν, η υλοποίηση αυτού του προγράμματος Matlab ήταν αναγκαία για να είναι δυνατή η χρήση των στατιστικών δελτίων του ΙΚΑ. Βάσει αυτών των στατιστικών δελτίων και με τη βοήθεια του επόμενου προγράμματος Matlab κατέστη δυνατή η παραγωγή δεδομένων που μεταβάλλονται ως προς το χρόνο.

Το δεύτερο πρόγραμμα που υλοποιήθηκε είναι το statist.m (Παράρτημα Γ.2). Ξεκινώντας από τα πρωτογενή δεδομένα του αρχείου test_december_07.xls και με βάση τα ποσοστά των εξαμηνιαίων στατιστικών του ΙΚΑ παράχθηκαν δεδομένα τόσο για προγενέστερα όσο και για μελλοντικά εξάμηνα (Ιούνιος 2004 έως Ιούνιος 2008). Έτσι έγινε δυνατή η παραγωγή δεδομένων οικονομικής δραστηριότητας που μεταβάλλονται με το χρόνο, βασικό στοιχείο των Αποθηκών Δεδομένων για εφαρμογές αναλυτικής επεξεργασίας (OLAP) δεδομένων και εξόρυξης γνώσης από πολυδιάστατες δομές.

Ο αλγόριθμος παραγωγής δεδομένων οικονομικής δραστηριότητας που μεταβάλλονται στον άξονα του χρόνου (πρόγραμμα statist.m) παρουσιάζεται με τη βοήθεια του ακόλουθου συνοπτικού παραδείγματος:

Τα εξαμηνιαία στατιστικά δελτία του ΙΚΑ (Παράρτημα Β.3) κατανέμουν τους αλλοδαπούς ενεργούς ασφαλισμένους σε τρεις μεγάλες κατηγορίες: Χώρες Ευρωπαϊκής Ένωσης (EU), Αλβανούς (AL) και Υπόλοιπους (ΟΤ). Το πρόγραμμα statist.m ελέγχει σε ποια κατηγορία-υπηκοότητα ανήκει ο κάθε ασφαλισμένος και χρησιμοποιεί τα αντίστοιχα στατιστικά αυτής της κατηγορίας. Για το συνοπτικό παράδειγμα θεωρούμε 5 οικονομικές δραστηριότητες (αντί των πραγματικών 17) και έστω ότι ο ασφαλισμένος ανήκει στην κατηγορία EU. Έστω ο παρακάτω πίνακας για δύο συνεχόμενα εξάμηνα της κατηγορίας EU:

Πίνακας 3-1. Παράδειγμα

Οικονομική Δραστηριότητα	Εξάμηνο Α (%)	Εξάμηνο Β (%)	Β-Α (%)
1	20	18	-2
2	15	16	1
3	10	8	-2
4	10	12	2
5	5	10	5

Θεωρούμε ότι το Εξάμηνο Α είναι το αρχικό και το Εξάμηνο Β είναι το τελικό, δηλαδή έχουμε διαθέσιμα στοιχεία για το Α και θέλουμε να παράξουμε στοιχεία για το Β. Έτσι με βάση τη στήλη Β-Α παρατηρούμε ότι οι οικονομικές δραστηριότητες 1 και 3 μειώνονται και οι υπόλοιπες αυξάνονται σε ποσοστό. Κάνουμε την εξής παραδοχή: *«Ασφαλισμένος που ανήκει σε οικονομική δραστηριότητα που αυξάνεται ή μένει σταθερή, δεν μετακινείται. Ασφαλισμένος που ανήκει σε οικονομική δραστηριότητα που μειώνεται μπορεί να μετακινηθεί μόνο σε οικονομική δραστηριότητα που αυξάνεται».* Άρα μετακίνηση μπορεί να γίνει μόνο για τους ασφαλισμένους που ανήκουν σε οικονομική δραστηριότητα που το ποσοστό της μειώνεται (οικονομικές δραστηριότητες 1,3 στο παράδειγμα). Για αυτούς τους ασφαλισμένους το πρόγραμμα κάνει δύο κληρώσεις με τη βοήθεια της συνάρτησης του Matlab rand() που παράγει τυχαίους αριθμούς με βάση την Ομοιόμορφη Κατανομή:

α) Κλήρωση πρώτη για το εάν ο ασφαλισμένος θα μετακινηθεί. Βλέπουμε στο παράδειγμα ότι στην οικονομική δραστηριότητα 1 σε 20 ασφαλισμένους μετακινούνται οι 2. Άρα το πρόγραμμα δίνει πιθανότητα 2/20=0,1 στον ασφαλισμένο αυτής της κατηγορίας να μετακινηθεί. Αντίστοιχα στην οικονομική δραστηριότητα 3 η πιθανότητα μετακίνησης είναι 2/10=0,2.

β) Κλήρωση δεύτερη για το προς ποια οικονομική δραστηριότητα θα μετακινηθεί ο ασφαλισμένος. Για όσους από την πρώτη κλήρωση αποφασίστηκε ότι θα μετακινηθούν, γίνεται αυτή η δεύτερη κλήρωση με την οποία θα αποφασιστεί σε ποια οικονομική δραστηριότητα που

αυξάνεται ως ποσοστό θα ενταχθούν τελικά. Άρα στο παράδειγμα ο ασφαλισμένος θα ενταχθεί σε κάποια οικονομική δραστηριότητα από τις 2,4,5. Από τη στήλη Β-Α προσθέτουμε τα ποσοστά αύξησης των γραμμών 2,4,5: Ptotal=1+2+5=8. Συνεπώς η μετακίνηση προς την οικονομική δραστηριότητα 2 θα γίνει με πιθανότητα P2=1/8, προς την 4 θα γίνει με πιθανότητα P4=2/8 και προς την στην 5 με P5=5/8. Σημειώνεται ότι προφανώς P2+P4+P5=1. Έτσι η οικονομική δραστηριότητα που αυξάνεται ποσοστιαία περισσότερο από τις άλλες (δηλαδή η 5) έχει και μεγαλύτερες πιθανότητες απορρόφησης ασφαλισμένων. Σημειώνεται ότι με την παραπάνω μέθοδο οι ποσοστιαίες μεταβολές στην οικονομική δραστηριότητα ανά εξάμηνο ακολουθούν τις πραγματικές ποσοστιαίες μεταβολές της οικονομικής δραστηριότητας όπως αυτές δίνονται από τους επίσημους στατιστικούς πίνακες του ΙΚΑ (Παράρτημα Β.3).

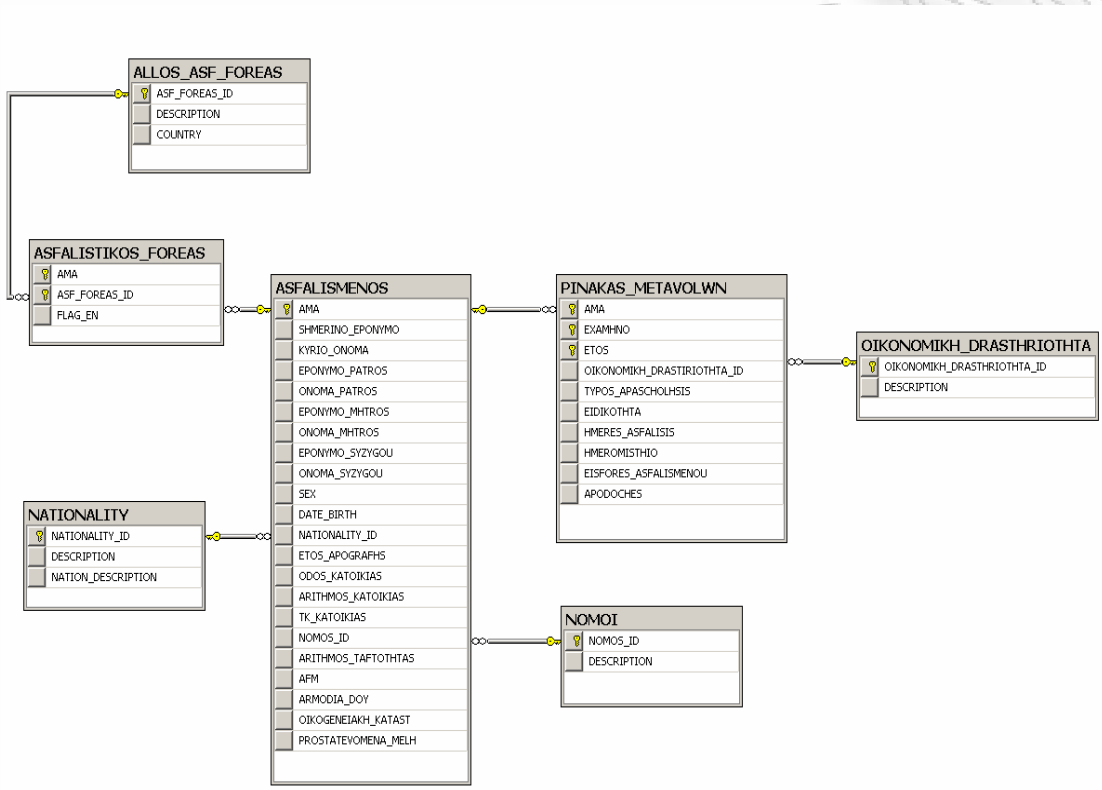
Το τελευταίο πρόγραμμα που υλοποιήθηκε είναι το apografh.m (Παράρτημα Γ.3). Αυτό το πρόγραμμα υπολογίζει το έτος απογραφής του κάθε ασφαλισμένου στο ασφαλιστικό ταμείο ΙΚΑ με βάση κάποιες παραδοχές που περιγράφονται στη συνέχεια. Καθώς δεν υπάρχουν επίσημα στατιστικά στοιχεία απογραφής του κάθε αλλοδαπού ασφαλισμένου, τόσο για το έτος προσέλευσής του στη χώρα όσο και για το έτος έναρξης ασφάλισης, επιλέχθηκαν οι παρακάτω παραδοχές που περίπου ανταποκρίνονται στα ποσοστά εισροής μεταναστών στη χώρα τις συγκεκριμένες χρονικές περιόδους. Σκοπός της όλης διαδικασίας είναι η προσθήκη ενός ακόμη πεδίου (Έτος απογραφής) στη βάση, το οποίο δίνει τη δυνατότητα υπολογισμού της παλαιότητας ενός ασφαλισμένου, δηλαδή του χρόνου ασφάλισής του στο συγκεκριμένο ασφαλιστικό φορέα. Στη συνέχεια δίνονται οι παραδοχές και περιγράφεται συνοπτικά το πρόγραμμα apografh.m.

Θεωρούμε ως ελάχιστη ηλικία απογραφής τα 17 έτη. Επίσης, θεωρούμε ότι όλες οι απογραφές γίνονται μεταξύ των ετών 1991 και 2004. Όσοι έχουν έτος γέννησης μεγαλύτερο ή ίσο του 1975 κατανέμονται ομοιόμορφα μεταξύ των ετών Έτος_Γέννησης+17 και 2004. Για τους υπόλοιπους ασφαλισμένους ελήφθησαν υπόψη και οι κατηγορίες EU, AL, OT, και ορίστηκαν τα ποσοστά κατανομής τους στις αντίστοιχες χρονικές περιόδους ως εξής:

EU: 45% ομοιόμορφα στο 1991-1997
55% ομοιόμορφα στο 1998-2004
AL: 70% ομοιόμορφα στο 1991-1996
20% ομοιόμορφα στο 1997-2001
10% ομοιόμορφα στο 2002-2004
OT**:** 40% ομοιόμορφα στο 1991-1997
60% ομοιόμορφα στο 1998-2004

3.2.3 Υλοποίηση της Βάσης Δεδομένων

Η βάση δεδομένων του ΙΚΑ-ΕΤΑΜ υλοποιήθηκε στο Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) του Microsoft SQL Server 2005 και τα δεδομένα, τόσο τα διαθέσιμα όσο και αυτά που παράχθηκαν, εισήχθησαν στους αντίστοιχους πίνακες με το εργαλείο Data Import for SQL Server 2005. Στη συνέχεια δίνεται το σχεσιακό σχήμα της βάσης δεδομένων του ΙΚΑ-ΕΤΑΜ ενώ τα SQL Scripts των πινάκων της βάσης δίνονται στο Παράρτημα Α.1.



Σχήμα 3-2. Το Σχεσιακό Σχήμα της Βάσης Δεδομένων του ΙΚΑ-ΕΤΑΜ

Ο σχεδιασμός και η υλοποίηση της παραπάνω βάσης δεδομένων στηρίχθηκε στις διαθέσιμες πρωτογενείς πηγές δεδομένων και στις απαιτήσεις και παραδοχές που καταγράφηκαν στην ενότητα 3.2.1. Ωστόσο, για την καλύτερη κατανόηση του περιεχομένου της βάσης και των παραδοχών βάσει των οποίων αυτή δημιουργήθηκε κρίνεται αναγκαία μια συνοπτική αναφορά στους πίνακες που την απαρτίζουν.

Πίνακας ASFALISMENOS: Στον πίνακα αυτό εισήχθησαν αυτούσια τα δεδομένα του αντιπροσωπευτικού δείγματος που δόθηκε σε αρχείο Excel, δηλαδή τα στοιχεία Αύξων Αριθμός που αντιστοιχίστηκε με το πεδίο του πρωτεύοντος κλειδιού του πίνακα ΑΜΑ – Αριθμός Μητρώου Ασφαλισμένου, Φύλο - SEX, Ημερομηνία Γέννησης – DATE_BIRTH, Εθνικότητα – NATIONALITY_ID και Νομός κατοικίας-εργασίας – NOMOS_ID. Επίσης, στο πεδίο ETOS_APOGRAFHS έγινε εισαγωγή του έτους απογραφής κάθε αλλοδαπού ασφαλισμένου, όπως αυτό παράχθηκε με το αντίστοιχο πρόγραμμα στο Matlab. Για τα υπόλοιπα πεδία του πίνακα, δηλαδή τα πεδία [SHMERINO_EPONYMO], [KYRIO_ONOMA], [EPONYMO_PATROS], [ONOMA_PATROS], [EPONYMO_MHTROS], [ONOMA_MHTROS], [EPONYMO_SYZYGOU], [ONOMA_SYZYGOU], [ODOS_KATOIKIAS], [AFM], [TK_ KATOIKIAS],[ARITHMOS_KATOIKIAS],[ARITHMOS_TAFTOTHTAS], [ARMODIA_DOY], [OIKOGENEIAKH_KATAST] και [PROSTATEVOMENA_MELH] δεν υπάρχουν διαθέσιμα στοιχεία και περιέχουν την τιμή NULL. Τα πεδία αυτά δημιουργήθηκαν ως υποδομή υποδοχής διαθέσιμων στοιχείων μελλοντικά, ενώ για τις ανάγκες της παρούσας εργασίας δεν θα

συμπεριλαμβάνονταν έστω και αν υπήρχε πρόσβαση σε αυτά καθώς αποτελούν ευαίσθητα προσωπικά δεδομένα.

Πίνακας NOMOI: Στον πίνακα αυτό περιλαμβάνονται 54 νομοί της Ελλάδας, δηλαδή οι 50 νομοί της χώρας πλην του νομού του Αγίου Όρους, ενώ ο νομός Αττικής διαιρείται στις νομαρχίες Ανατολικής και Δυτικής Αττικής, Πειραιά και Αθηνών. Επίσης, στον πίνακα αυτό διατηρήθηκε ως πρωτεύον κλειδί ο κωδικός του νομού όπως αυτός ήταν διαθέσιμος στο αρχικό αρχείο Excel.

Πίνακας NATIONALITY: Στο αρχικό αρχείο Excel η εθνικότητα του ασφαλισμένου δίνεται με την επίσημη συντομογραφία κάθε χώρας με δύο λατινικούς χαρακτήρες. Στον πίνακα αυτό ορίστηκε ως πρωτεύον κλειδί η συντομογραφία που αντιστοιχεί σε κάθε χώρα και τελικά συμπεριλήφθηκαν 241 χώρες, σύμφωνα με τη επίσημη λίστα χωρών που δίνεται από τον Οργανισμό Ηνωμένων Εθνών. Παράλληλα, ο πίνακας αυτός περιλαμβάνει την πλήρη περιγραφή της ονομασίας κάθε χώρας, ενώ το πεδίο NATION_DESCRIPTION διατηρεί το βασικό διαχωρισμό των αλλοδαπών ασφαλισμένων του ΙΚΑ σε αλλοδαπούς που προέρχονται από χώρες της ΕΥΡΩΠΑΪΚΗΣ ΈΝΩΣΗΣ, την ΑΛΒΑΝΙΑ ή από ΆΛΛΗ ΧΩΡΑ.

Πίνακες ASFALISTIKOS_FOREAS & ALLOS_ASF_FOREAS: Οι πίνακες αυτοί δεν περιέχουν στοιχεία και δημιουργήθηκαν ως υποδομή για μελλοντική χρήση με βάση τις απαιτήσεις που καταγράφηκαν στην παράγραφο 3.2.1. Ο ρόλος των πινάκων αυτών είναι η διατήρηση του Ιστορικού Ασφάλισης, δηλαδή της ασφαλιστικής ιστορίας κάθε ασφαλισμένου, καθώς η φύση της απασχόλησης του εργαζόμενου καθορίζει και τον φορέα ασφάλισής του. Κάθε ασφαλισμένος κατά τη διάρκεια της παραγωγικής του ηλικίας μπορεί να εναλλάσσεται ανάμεσα στην απασχόληση με εξαρτημένη εργασία ή στην απασχόληση ως ελεύθερος επαγγελματίας και παράλληλα να ασφαλίζεται σε διαφορετικούς ασφαλιστικούς φορείς, ανάλογα και με τη χώρα που κάθε φορά δραστηριοποιείται.

Πίνακας PINAKAS_METAVOLWN: Ο πίνακας αυτός διαθέτει ένα τριπλό πρωτεύων κλειδί που αποτελείται από τα πεδία ΑΜΑ, ΕΧΑΜΗΝΟ και ΕΤΟS. Ουσιαστικά συνδέει κάθε ασφαλισμένο με την οικονομική δραστηριότητά του όπως αυτή διαμορφώθηκε για την περίοδο Ιουνίου 2004 έως Ιουνίου 2008 (εννέα εξάμηνα) με βάση τα επίσημα εξαμηνιαία στατιστικά δελτία του ΙΚΑ (Παράρτημα Β.3) με την παραδοχή βέβαια ότι κάθε ασφαλισμένος απασχολείται σε μία και μόνο μία οικονομική δραστηριότητα. Η αρχική επιδίωξη ήταν να σχεδιαστεί ένα πίνακας που θα παρέχει οικονομικές πληροφορίες για κάθε ασφαλισμένο και για κάθε περίοδο οικονομικής δραστηριοποίησής του. Σύμφωνα με τον Οδηγό Σύνδεσης Κωδικών (Ο.ΣΥ.Κ) του ΙΚΑ, από την οικονομική δραστηριότητα κάθε ασφαλισμένου απορρέουν ο τύπος απασχόλησης, η ειδικότητα, το ημερομίσθιο, οι εισφορές του ασφαλισμένου και τελικά οι νόμιμες αποδοχές του. Ωστόσο, στα πλαίσια υλοποίησης της παρούσας εργασίας δεν υπήρχε πρόσβαση στα αντίστοιχα στοιχεία που αποτελούν δεδομένα των Αναλυτικών Περιοδικών Δηλώσεων (ΑΠΔ) που καταρτίζουν και καταθέτουν οι εργοδότες στα τμήματα Εσόδων των Υποκαταστημάτων του ΙΚΑ. Επομένως, τα πεδία του πίνακα [TYPOS_APASCHOLHSHS], [EIDIKOTHTA], HMERES_ASFALISHS], [HMEROMISTHIO], [EISFORES_ASFALISMENOY],[APODOCHES] είναι κενά, δηλαδή περιέχουν την τιμή NULL.

Πίνακας ΟΙΚΟΝΟΜΙΚΗ_DRASTHRIOTHTA: Ο πίνακας αυτός περιέχει την περιγραφή κάθε μιας από τις 17 συνολικά κατηγορίες οικονομικής δραστηριότητας όπως αυτές δίνονται στο ΣΤΑΚΟΔ 2003 (Παράρτημα Β.2 - Στατιστική Ταξινόμηση των Κλάδων της Οικονομικής Δραστηριότητας) και χρησιμοποιούνται στα εξαμηνιαία στατιστικά δελτία του ΙΚΑ (Παράρτημα Β.3).

3.3 Η Αποθήκη των Μεταναστευτικών Δεδομένων (ΙΚΑ_ΟΑΕΕ_DW)

3.3.1 Κατανόηση του Προβλήματος - Επιλογή Δεδομένων - Ορισμός Διαστάσεων

Η Αποθήκη υλοποιήθηκε με τέτοιο τρόπο ώστε να συγκεντρωθούν σε αυτή τα διαθέσιμα ετερογενή δημογραφικά, οικονομικά και γεωγραφικά δεδομένα των ενεργών αλλοδαπών ασφαλισμένων των δύο μεγαλύτερων ασφαλιστικών οργανισμών της χώρας, του ΙΚΑ-ΕΤΑΜ και του ΟΑΕΕ με διαδικασίες που εξασφάλισαν τη συνέπεια δεδομένων, οι οποίες θα αναλυθούν

στην επόμενη ενότητα. Η Αποθήκη Δεδομένων σχεδιάστηκε ως σχεσιακή βάση που όμως θα εξυπηρετεί τους παρακάτω στόχους:

1. Να παρέχει τη δυνατότητα δημιουργίας οποιουδήποτε είδους πολυδιάστατης βάσης δεδομένων πάνω από αυτή.
2. Να παρέχει τη δυνατότητα αναλυτικής επεξεργασίας των δεδομένων με τεχνικές OLAP (Online Analytical Processing).
3. Να παρέχει τη δυνατότητα εφαρμογής τεχνικών Εξόρυξης Γνώσης (Data Mining).
4. Να παρέχει τη δυνατότητα επαναφόρτωσης των πολυδιάστατων κύβων χωρίς πρόσβαση στα πρωτογενή δεδομένα.
5. Γενικά, να παρέχει τη δυνατότητα πραγματοποίησης πολυδιάστατων αναλύσεων με βάση τις θεματικές περιοχές που έχουν προσδιοριστεί από τις αντίστοιχες σχετικές απαιτήσεις.

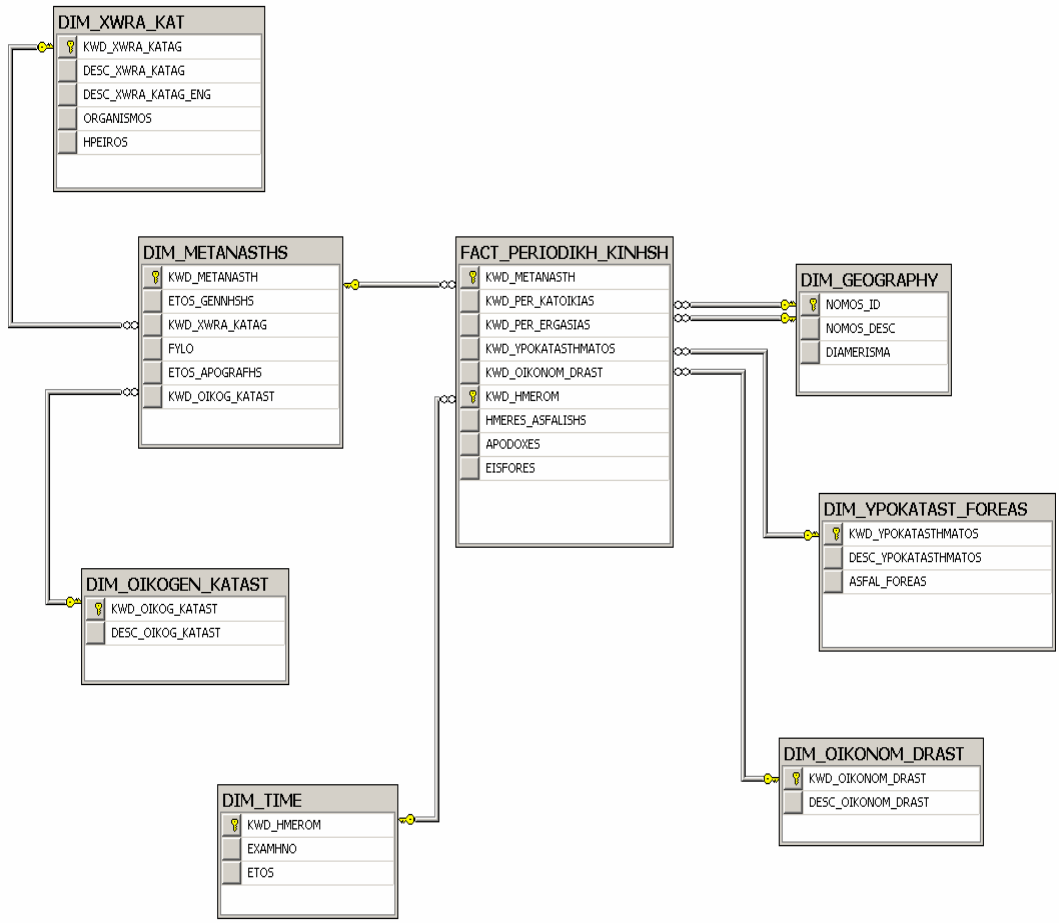
Οι παραπάνω στόχοι και οι διαθέσιμες πρωτογενείς πηγές δεδομένων καθόρισαν τη λογική και φυσική σχεδίαση της Αποθήκης Δεδομένων, δηλαδή τους πίνακες διαστάσεων (dimension tables), τους πίνακες γεγονότων (fact tables) και τα μετρήσιμα μεγέθη (measures), τα οποία θα αποτελέσουν και τα αντικείμενα περαιτέρω ανάλυσης. Κατά το λογικό και φυσικό σχεδιασμό της Αποθήκης και την πολυδιάστατη μοντελοποίηση των δεδομένων, καταρτίστηκε ένα σύνολο μετρήσιμων μεγεθών – τιμών, οι οποίες εξαρτώνται από ένα σύνολο διαστάσεων και ιεραρχιών. Οι διαστάσεις και οι ιεραρχίες χαρακτηρίζουν μοναδικά τις μετρήσιμες τιμές και τις τοποθετούν σε μια περιοχή του πολυδιάστατου χώρου. Για τις ανάγκες της εργασίας, Θα μπορούσαν να καθοριστούν γενικά οι παρακάτω διαστάσεις και με τις ιεραρχίες τους:

1) Χρόνος
   i. Έτος
   ii. Εξάμηνο
2) Δημογραφικά Στοιχεία
   i. Έτος Γέννησης
   ii. Χώρα Καταγωγής
   iii. Φύλο
   iv. Τόπος Κατοικίας
   v. Οικογενειακή Κατάσταση (Άγαμος, Έγγαμος, Διαζευγμένος, Χήρος)
3) Γεωγραφική Περιοχή
   i. Ήπειρος
   ii. Οργανισμός
   iii. Χώρα
   iv. Γεωγραφικό Διαμέρισμα
   v. Νομός
4) Κατηγορίες Οικονομικής Δραστηριότητας (π.χ. αλιεία, γεωργία – κτηνοτροφία – θήρα - δασοκομία, ορυχεία και λατομεία, μεταποιητικές βιομηχανίες κλπ).
5) Φορέας Ασφάλισης
   i. ΙΚΑ-ΕΤΑΜ
   ii. ΟΑΕΕ
6) Έτος Απογραφής άμεσα ασφαλισμένου.
7) Εύρος εξαμηνιαίων εισφορών ασφαλισμένου.
8) Εύρος ετήσιων αποδοχών ασφαλισμένου.
9) Ημέρες ασφάλισης ανά μήνα και έτος.

Ωστόσο, λόγω της ανομοιογένειας των διαθέσιμων πρωτογενών δεδομένων κατά την υλοποίηση της Αποθήκης Δεδομένων δεν διατηρήθηκαν αυστηρά οι διαστάσεις και οι αντίστοιχες ιεραρχίες τους, όπως ορίστηκαν προηγουμένως. Τα διαθέσιμα δεδομένα καθόρισαν ουσιαστικά τη σχεδίαση της Αποθήκης, ενώ παράλληλα δημιουργήθηκαν και πεδία ως υποδομή για υποδοχή δεδομένων που θα είναι μελλοντικά διαθέσιμα.

3.3.2 Υλοποίηση της Αποθήκης Δεδομένων

Σκοπός της παρούσας ενότητας είναι η παρουσίαση της Αποθήκης Μεταναστευτικών Δεδομένων που υλοποιήθηκε μετά από τον συγκερασμό των ετερογενών πηγών πρωτογενών δεδομένων που περιγράφτηκαν στις προηγούμενες ενότητες καθώς και η ανάλυση των διαδικασιών εισαγωγής δεδομένων που ακολουθήθηκαν. Για την υλοποίηση της Αποθήκης χρησιμοποιήθηκε και πάλι το Management Studio του Microsoft SQL Server 2005. Στη συνέχεια δίνεται το σχεσιακό σχήμα της Αποθήκης Δεδομένων, Σχήμα 3-3.



Σχήμα 3-3. Σχεσιακό σχήμα της Αποθήκης Δεδομένων ΙΚΑ_ΟΑΕΕ_DW

Η μοντελοποίηση της Αποθήκης Δεδομένων έγινε με συνδυασμό σχήματος αστέρα (star schema - κεντρικός πίνακας FACT_PERIODIKH_KINHSH) και σχήματος χιονονιφάδας (snowflake schema - με κανονικοποίηση της διάστασης DIM_METANASTHS στους πίνακες χώρα καταγωγής – DIM_XWRA_KAT και οικογενειακή κατάσταση –DIM_OIKOGEN_KATAST). Τα SQL Scripts δημιουργίας των πινάκων της Αποθήκης ΙΚΑ_ΟΑΕΕ_DW δίνονται στο Παράρτημα Α.2. Στο Παράρτημα Α.3 δίνονται τα SQL Scripts που τροφοδότησαν τους αντίστοιχους πίνακες της Αποθήκης Δεδομένων με δεδομένα από τη σχεσιακή βάση του ΙΚΑ-ΕΤΑΜ. Παράλληλα σε κάποιες εισαγωγές «φορτώθηκαν» δεδομένα του ΙΚΑ και του ΟΑΕΕ/ΤΕΒΕ ταυτόχρονα ενώ τα SQL Scripts που τροφοδότησαν την Αποθήκη Δεδομένων μόνο με στοιχεία του ΟΑΕΕ/ΤΕΒΕ δεν αποτελούν αντικείμενο της παρούσας εργασίας.

Ωστόσο, για την καλύτερη κατανόηση του περιεχομένου της Αποθήκης και των παραδοχών βάσει των οποίων τα δεδομένα «φορτώθηκαν» σε αυτή, ώστε να εξασφαλιστεί η συνέπεια των δεδομένων, ακολουθεί μια συνοπτική αναφορά στους πίνακες που την απαρτίζουν.

Πίνακας FACT_PERIODIKH_KINHSH: Τα πεδία του συγκεκριμένου πίνακα αποτελούνται από τα κλειδιά προς τους σχετιζόμενους πίνακες διαστάσεων, [KWD_METANASTH], [KWD_PER_KATOIKIAS], [KWD_PER_ERGASIAS], [KWD_YPOKATASTHMATOS], [KWD_ OIKONOM_DRAST], [KWD_HMEROM], και τα μέτρα (κάποια μετρήσιμα μεγέθη) [HMERES_ASFALISHS], [APODOXES], [EISFORES]. Από τα μέτρα, τα δύο πρώτα δημιουργήθηκαν ως υποδομή για μελλοντική εισαγωγή δεδομένων, ενώ το πεδίο [EISFORES] περιέχει στοιχεία εισφορών μόνο για τους ασφαλισμένους του ΟΑΕΕ και μόνο για την περίοδο των ετών 2007-2009. Τα δεδομένα του πίνακα «φορτώθηκαν» από τις βάσεις δεδομένων του ΙΚΑ και του ΟΑΕΕ/ΤΕΒΕ. Τα δεδομένα του ΙΚΑ προήλθαν από τους πίνακες ASFALISMENOS και PINAKAS_METAVOLWN. Επειδή, στη βάση του ΙΚΑ δεν κρατούνται δεδομένα για τα υποκαταστήματα του ασφαλιστικού φορέα, σε αντίθεση με τη βάση του ΟΑΕΕ, δόθηκε ο κωδικός '0' ως [KWD_YPOKATASTHMATOS] του ΙΚΑ. Τέλος, το σύνολο των αλλοδαπών ασφαλισμένων του ΙΚΑ (περίπου 65500 εγγραφές) εξετάζεται ως προς τις μεταβολές της οικονομικής τους δραστηριότητας για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 (εννέα εξάμηνα) γεγονός που αυξάνει κατά πολύ το μέγεθος του πίνακα, το οποίο ανέρχεται μαζί με τους ασφαλισμένους του ΟΑΕΕ σε 941428 εγγραφές. Λόγω του μεγάλου μεγέθους του συγκεκριμένου πίνακα δημιουργήθηκαν ευρετήρια (indexes) στα πεδία των ξένων κλειδιών, δηλαδή [KWD_PER_KATOIKIAS], [KWD_PER_ERGASIAS], [KWD_YPOKATASTHMATOS] και [KWD_OIKONOM_DRAST], ώστε να επιτυγχάνεται καλύτερη απόδοση όταν εκτελούνται αναζητήσεις.

Πίνακας DIM_METANASTHS: Τόσο οι ασφαλισμένοι του ΙΚΑ όσο και οι Ασφαλισμένοι του ΟΑΕΕ, όταν απογράφονται στον ασφαλιστικό φορέα τους λαμβάνουν ένα μοναδικό Αριθμό Μητρώου Ασφαλισμένου (ΑΜΑ). Επίσης, και οι δύο ασφαλιστικοί οργανισμοί διατηρούν στοιχεία για το έτος γέννησης, το φύλο, τη χώρα καταγωγής, την οικογενειακή κατάσταση και το έτος απογραφής των ασφαλισμένων τους. Στην προκειμένη περίπτωση, τα αντίστοιχα δεδομένα εισήχθησαν από τις βάσεις στον συγκεκριμένο πίνακα της Αποθήκης, με τη μόνη διαφορά ότι για τους ασφαλισμένους του ΙΚΑ δεν υπάρχουν διαθέσιμα στοιχεία οικογενειακής κατάστασης. Ουσιαστικά, ο πίνακας αυτός συγκεντρώνει δημογραφικά δεδομένα για τους αλλοδαπούς απασχολούμενους των δύο ασφαλιστικών οργανισμών, και αποτελεί τον δεύτερο σε σειρά από άποψη μεγέθους πίνακα της Αποθήκης με συνολικά 136146 εγγραφές. Επομένως, στα πλαίσια επίτευξης καλύτερης απόδοσης δημιουργήθηκαν ευρετήρια πάνω στα πεδία έτος γέννησης, έτος απογραφής και φύλο.

Πίνακας DIM_XWRA_KAT: Στον πίνακα αυτό, που ουσιαστικά κανονικοποιεί τον πίνακα DIM_METANASTHS, συγκεντρώνονται οι 241 χώρες που ανήκουν στην επίσημη λίστα των χωρών που δίνεται από τον Οργανισμό Ηνωμένων Εθνών. Ο κωδικός των χωρών είναι η επίσημη συντομογραφία τους με δύο λατινικούς χαρακτήρες, ενώ στα υπόλοιπα πεδία οι χώρες δίνονται με την πλήρη ονομασία τους στα ελληνικά και στα αγγλικά και παράλληλα διαχωρίζονται σε Οργανισμούς (π.χ. Ευρωπαϊκή Ένωση κλπ) και Ηπείρους, ώστε να επιτευχθεί ο συγκερασμός της διαθέσιμης πληροφορίας που υπάρχει στους δύο ασφαλιστικούς οργανισμούς.

Πίνακας DIM_OIKOGEN_KATAST: Στον πίνακα αυτό συγκεντρώνεται διαθέσιμη πληροφορία της βάσης του ΟΑΕΕ, γι'αυτό και στο Παράρτημα Α.3 δεν υπάρχει SQL Script εισαγωγής δεδομένων, ενώ παράλληλα γίνεται κανονικοποίηση του πίνακα DIM_METANASTHS.

Πίνακας DIM_TIME: Η διάσταση του χρόνου είναι βασικό δομικό στοιχείο κάθε αποθήκης δεδομένων και στη συγκεκριμένη περίπτωση ο χρόνος αναλύεται σε δώδεκα (KWD_HMEROM 1 έως 12) εξάμηνα ($1^{ο}$ και $2^{ο}$) για την περίοδο των ετών 2004 έως 2009.

Πίνακας DIM_OIKONOM_DRAST: Στον πίνακα αυτό συγκεντρώνεται η πληροφορία που παράχθηκε για τις μεταβολές της οικονομικής δραστηριότητας των ασφαλισμένων του ΙΚΑ, σύμφωνα με επίσημα δημοσιευμένα στατιστικά δελτία του ΙΚΑ-ΕΤΑΜ, δηλαδή 17 συνολικά κατηγορίες σύμφωνα με το ΣΤΑΚΟΔ 2003. Για τους ασφαλισμένους του ΟΑΕΕ, δεν υπάρχουν αντίστοιχα αξιοποιήσιμα δημοσιευμένα στοιχεία καθώς ο ασφαλιστικός φορέας ακολουθεί

τελείως διαφορετική κωδικοποίηση για την αποτύπωση της οικονομικής δραστηριότητας των ασφαλισμένων του. .

Πίνακας DIM_GEOGRAPHY: Στον πίνακα αυτό διατηρήθηκε αυτούσια η διαθέσιμη πληροφορία του πίνακα ΝΟΜΟΙ της βάσης δεδομένων του ΙΚΑ και πάνω σε αυτή ενσωματώθηκαν τα αντίστοιχα στοιχεία του ΟΑΕΕ. Δηλαδή, ο πίνακας περιλαμβάνει τους 54 νομούς της Ελλάδας, δηλαδή τους 50 νομούς της χώρας πλην του νομού του Αγίου Όρους, με παράλληλη διαίρεση του νομού Αττικής στις νομαρχίες Ανατολικής και Δυτικής Αττικής, Πειραιά και Αθηνών. Επίσης, προστέθηκε ένα επιπλέον πεδίο, το [DIAMERISMA], ώστε κάθε νομός να αντιστοιχίζεται στο γεωγραφικό του διαμέρισμα. Τέλος, η σύνδεση της συγκεκριμένης διάστασης με τον κεντρικό πίνακα γεγονότων (FACT_PERIODIKH_KINHSH) έγινε με την παραδοχή ότι ο κάθε αλλοδαπός ασφαλισμένος στο ΙΚΑ ή στον ΟΑΕΕ διαμένει και απασχολείται στον ίδιο νομό, δηλαδή η περιοχή κατοικίας και εργασίας συμπίπτουν.

Πίνακας DIM_YPOKATAST_FOREAS: Ο πίνακας αυτός δημιουργήθηκε κυρίως από την ανάγκη διαχωρισμού των ασφαλισμένων, σε ασφαλισμένους του ΙΚΑ-ΕΤΑΜ και σε ασφαλισμένους του ΟΑΕΕ, σε κάποιο στάδιο της περαιτέρω ανάλυσης. Καθώς η Αποθήκη συγκεντρώνει διαθέσιμη πληροφορία για δύο διαφορετικούς πληθυσμούς ασφαλισμένων, ασφαλισμένους με εξαρτημένη εργασία και ελεύθερους επαγγελματίες, κρίθηκε σκόπιμο να παρέχεται παράλληλα με τη δυνατότητα ανάλυσης των δεδομένων στο σύνολο και η δυνατότητα ανάλυσης των δεδομένων ανά ασφαλιστικό φορέα. Επίσης, στον πίνακα αυτό έχουν εισαχθεί και τα υποκαταστήματα του ΟΑΕΕ ώστε να είναι εφικτή η ανάλυση των δεδομένων σε μεγαλύτερο επίπεδο λεπτομέρειας – επόμενο επίπεδο στην ιεραρχία, δηλαδή ανάλυση δεδομένων ανά υποκατάστημα του Οργανισμού Ασφάλισης Ελευθέρων Επαγγελματιών.

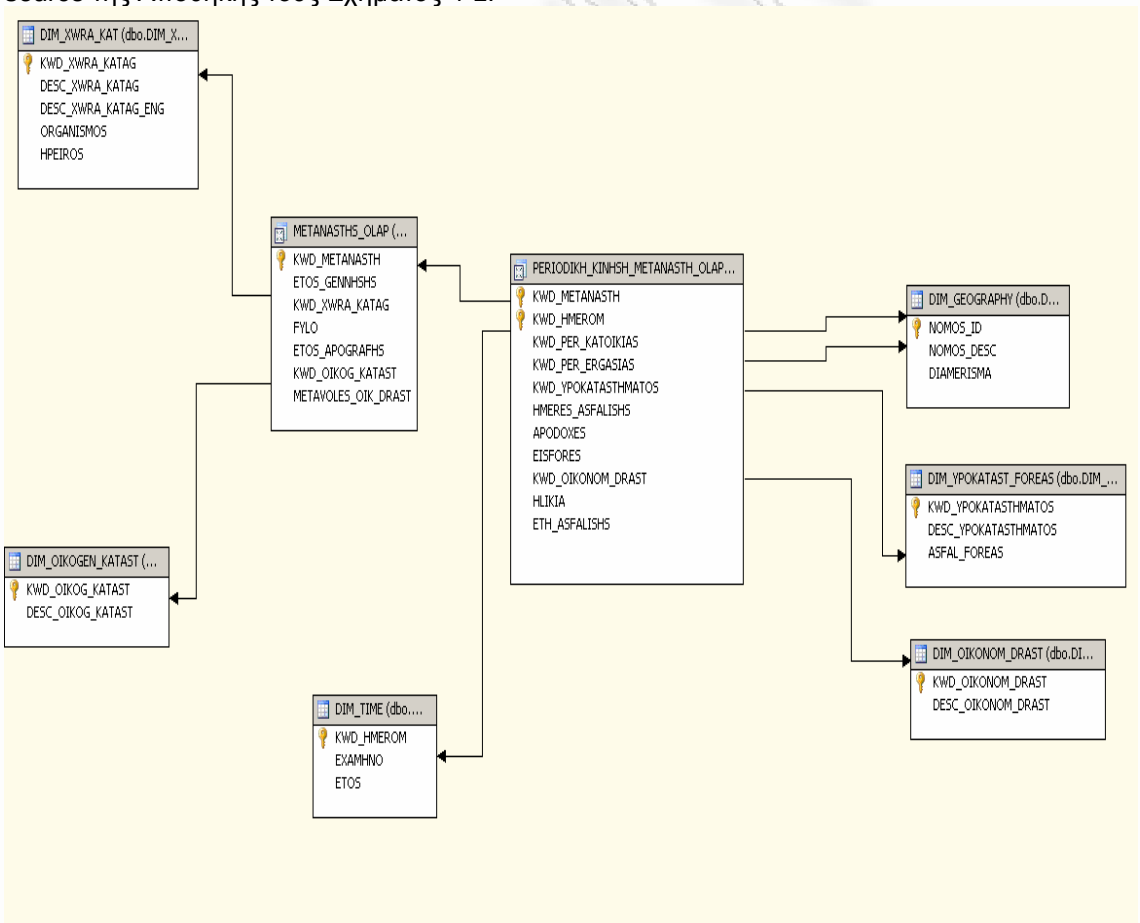
Συνοψίζοντας, η υλοποίηση της Αποθήκης Δεδομένων πραγματοποιήθηκε με αρκετές παραδοχές ως προς την εισαγωγή των δεδομένων ώστε να εξασφαλιστεί η συνέπειά τους. Όσον αφορά στα δεδομένα του ΙΚΑ «φορτώθηκαν» στην Αποθήκη σχεδόν αυτούσια και δεν εφαρμόστηκαν ιδιαίτερες διαδικασίες καθαρισμού και μετασχηματισμού των δεδομένων, αφού προήλθαν από μια βάση δεδομένων της οποίας ο μεγαλύτερος όγκος πληροφορίας παράχθηκε ειδικά για τις ανάγκες ολοκλήρωσης της παρούσας εργασίας. Ωστόσο, στη συγκεκριμένη Αποθήκη υπάρχουν και πεδία με κενές τιμές, μερικά εκ των οποίων δημιουργήθηκαν ως υποδομή για εισαγωγή στοιχείων που θα είναι μελλοντικά διαθέσιμα, τα περισσότερα όμως είναι αποτέλεσμα της τεράστιας ανομοιογένειας των διαθέσιμων δεδομένων των δύο πληθυσμών ασφαλισμένων και κατ΄επέκταση της αδυναμίας πλήρους πρόσβασης στα πρωτογενή δεδομένα των δύο ασφαλιστικών οργανισμών. Οι αδυναμίες της συγκεκριμένης Αποθήκης Μεταναστευτικών Δεδομένων δημιούργησαν προβλήματα κατά την OLAP ανάλυση και την εφαρμογή τεχνικών εξόρυξης γνώσης, θέματα που αναπτύσσονται αναλυτικά στα επόμενα κεφάλαια, τα οποία όπως θα δούμε ξεπεράστηκαν με τη δημιουργία όψεων (views) που χρησιμοποιήθηκαν ως πίνακες στους κύβους που υλοποιήθηκαν. Επομένως, τα πολυδιάστατα μοντέλα δεδομένων μπορούν να δημιουργηθούν χρησιμοποιώντας τόσο σχεσιακούς πίνακες μιας βάσης δεδομένων όσο και βοηθητικούς πίνακες, καθώς σε κάθε επίπεδο της ανάλυσης αυτό που ενδιαφέρει είναι η καλύτερη παρουσίαση της εξαγόμενης πληροφορίας.

4. Αναλυτική Επεξεργασία Δεδομένων (OLAP Ανάλυση)

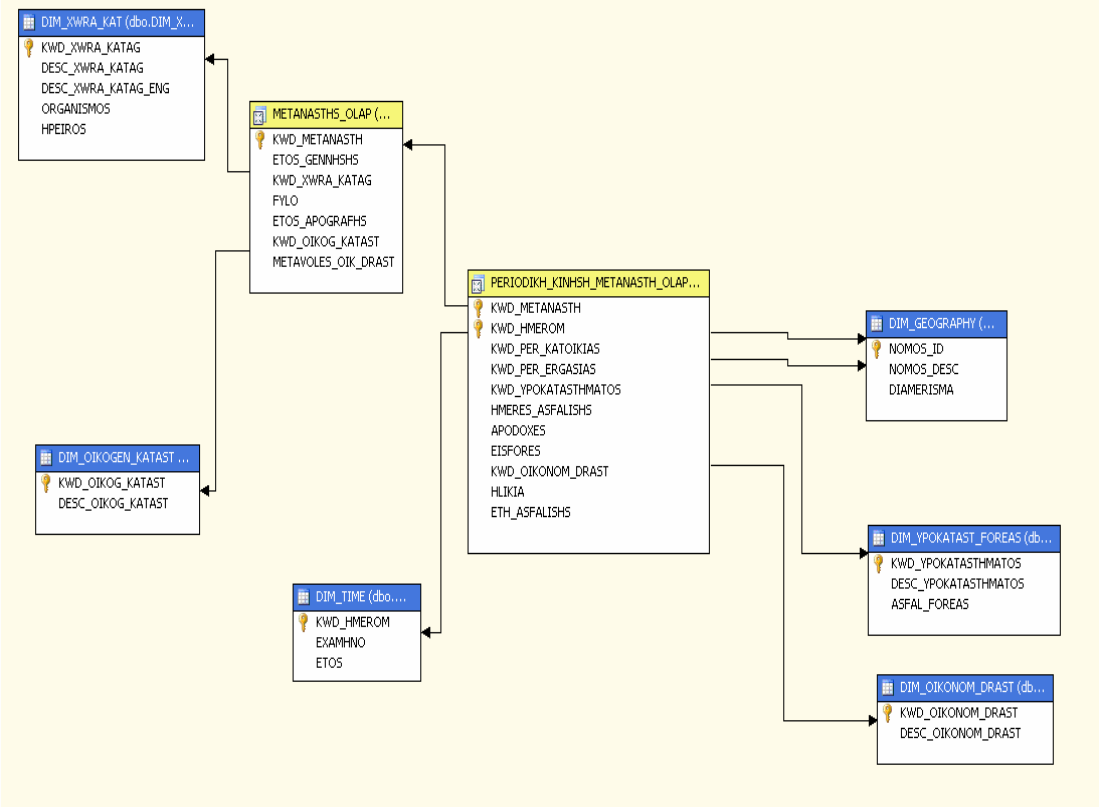
Σκοπός της παρούσας ενότητας είναι η παρουσίαση των εξαγόμενων αποτελεσμάτων και συμπερασμάτων που προήλθαν από την αναλυτική επεξεργασία των δεδομένων της Αποθήκης Μεταναστευτικών Δεδομένων στο περιβάλλον του Business Intelligence Management Studio του Microsoft SQL Server 2005 Analysis Services. Ωστόσο, λόγω της ιδιαιτερότητας που παρουσιάζουν τα διαθέσιμα δεδομένα των δύο ασφαλιστικών οργανισμών επιλέχθηκε μια πιο ευέλικτη μορφή (data source) της αρχικής Αποθήκης Δεδομένων με κατάλληλη προπαρασκευή των δεδομένων. Επομένως, το πολυδιάστατο μοντέλο δεδομένων – κύβος που δημιουργήθηκε για την OLAP ανάλυση, με τις αντίστοιχες διαστάσεις, ιεραρχίες και μέτρα, βασίστηκε τόσο στους αρχικούς πίνακες του σχεσιακού σχήματος της Παραγράφου 3.3.2 όσο και σε βοηθητικούς πίνακες – όψεις (views) που χρησιμοποιήθηκαν με στόχο την εξαγωγή καλύτερων αποτελεσμάτων.

4.1 Προπαρασκευή της Αποθήκης Δεδομένων για την OLAP Ανάλυση

Για την πραγματοποίηση OLAP Ανάλυσης σε κύβο με διαστάσεις, ιεραρχίες και μέτρα, δημιουργήθηκε μια πιο ευέλικτη μορφή πηγής δεδομένων (data source) από αυτή της αρχικής Αποθήκης Μεταναστευτικών Δεδομένων, της οποίας η σχεσιακή αναπαράσταση οποία δίνεται στο παρακάτω Σχήμα 4-1. Στο Σχήμα 4-2 δίνεται ο κύβος που δημιουργήθηκε πάνω στο data source της Αποθήκης τους Σχήματος 4-1.



Σχήμα 4-1. Η Αποθήκη Δεδομένων IKA_OAEE_DW.ds, που θα χρησιμοποιηθεί ως data source στον κύβο IKA_OAEE_DW.cube



Σχήμα 4-2. Ο κύβος IKA_OAEE_DW.cube

Το Σχήμα 4-1 της Αποθήκης Δεδομένων διατηρεί την ίδια δομή με το σχεσιακό σχήμα της Παραγράφου 3.3.2, είναι δηλαδή Σχήμα Χιονονιφάδας (Snowflake Schema). Η κύρια διαφορά του Σχήματος 4-1 με το αρχικό σχεσιακό σχήμα είναι οι βοηθητικοί πίνακες PERIODIKH_KINHSH_METANASTH_OLAP και METANASTHS_OLAP που αντικατέστησαν τους πίνακες FACT_PERIODIKH_KINHSH και DIM_METANASTHS αντίστοιχα. Η αντικατάσταση πραγματοποιήθηκε για τη δημιουργία επιπλέον πεδίων, χρήσιμων για τους σκοπούς της OLAP ανάλυσης.

Ο πίνακας γεγονότων PERIODIKH_KINHSH_METANASTH_OLAP αποτελεί μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα FACT_PERIODIKH_KINHSH, ενώ παράλληλα διαθέτει δύο επιπλέον πεδία – μέτρα, τα [HLIKIA] και [ETH_ASFALISHS]. Στα πεδία αυτά, τόσο η ηλικία όσο και τα έτη ασφάλισης, δηλαδή ο χρόνος ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα, υπολογίζονται για τις χρονικές περιόδους που καλύπτει η Αποθήκη Δεδομένων, δηλαδή για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 για τους ασφαλισμένους του IKA_ETAM και για την περίοδο Ιούνιος 2007 έως Δεκέμβριος 2009 για τους ασφαλισμένους του ΟΑΕΕ. Το SQL Script δημιουργίας του view PERIODIKH_KINHSH_METANASTH_OLAP δίνεται στο Παράρτημα Α.4.

Ο πίνακας METANASTHS_OLAP, που στον κύβο χρησιμοποιείται ως πίνακας διάστασης αλλά και ως δευτερεύον πίνακας γεγονότων, αποτελεί επίσης μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα DIM_METANASTHS, ενώ διαθέτει ένα επιπλέον πεδίο το [METAVOLES_OIK_DRAST]. Το πεδίο αυτό δημιουργήθηκε από την ανάγκη για καταγραφή πληροφορίας που σχετίζεται με τη μεταβολή της οικονομικής δραστηριότητας των ασφαλισμένων ως προς τις χρονικές περιόδους που εξετάζονται. Εξ' ορισμού, αλλά και λόγω της έλλειψης διαθέσιμων στοιχείων, για τους ασφαλισμένους του ΟΑΕΕ θεωρήθηκε ότι διατηρούν σταθερή την οικονομική τους δραστηριότητα, καθώς απασχολούνται ως ελεύθεροι επαγγελματίες. Όσον αφορά στους ασφαλισμένους του ΙΚΑ-ΕΤΑΜ, για τους οποίους παράχθηκαν στοιχεία οικονομικής δραστηριότητας για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 σύμφωνα με τα επίσημα δημοσιευμένα εξαμηνιαία στατιστικά δελτία του ΙΚΑ, το συγκεκριμένο πεδίο καταγράφει πληροφορία σχετικά με το αν κάποιοι ασφαλισμένοι διατηρούν την οικονομική τους δραστηριότητα για ορισμένες χρονικές περιόδους ή αλλάζουν συνεχώς. Με

αυτό τον τρόπο κατέστη δυνατή η παρακολούθηση της κινητικότητας των ασφαλισμένων του ΙΚΑ ανάμεσα στις διάφορες κατηγορίες οικονομικής δραστηριότητας, καθώς απασχολούνται με εξαρτημένη εργασία και κατά τη διάρκεια της ασφαλιστικής τους ιστορίας μπορούν να εναλλάσσουν οικονομικές δραστηριότητες και κατ' επέκταση εργοδότες. Στην περίπτωση που υπήρχαν διαθέσιμα στοιχεία μητρώου εργοδοτών θα ήταν πιο εύκολη η εξαγωγή συμπερασμάτων σχετικά με την εργασιακή κινητικότητα των ασφαλισμένων του ΙΚΑ. Το SQL Script δημιουργίας του view METANASTHS_OLAP δίνεται στο Παράρτημα Α.4.

Το Σχήμα 4-2 απεικονίζει τον κύβο που δημιουργήθηκε πάνω στο data source του Σχήματος 4-1, ο οποίος αποτελείται από δύο πίνακες γεγονότων (PERIODIKH _KINHSH_METANASTH_OLAP και METANASTHS_OLAP) με τα αντίστοιχα μέτρα και τις διαστάσεις τους. Στον κύβο έχουν διατηρηθεί και κάποια πεδία της αρχικής Αποθήκης Δεδομένων με κενές τιμές που ωστόσο δεν επηρεάζουν σημαντικά την OLAP ανάλυση.

Στη συνέχεια παραθέτονται τρεις πίνακες οι οποίοι συνοψίζουν τις διαστάσεις, ιεραρχίες και τα μέτρα του κύβου ΙΚΑ_ΟΑΕΕ_DW.cube. Ο πίνακας 4-1 παρουσιάζει όλες τις διαστάσεις του κύβου με τις αντίστοιχες ιεραρχίες τους. Στον πίνακα 4-2 περιγράφονται οι ομάδες μέτρων (measure groups), δηλαδή τα μετρήσιμα μεγέθη του κύβου, που δημιουργούνται και επιλέγονται με τον Cube Wizard. Τέλος, στον πίνακα 4-3 παρουσιάζονται τα επιπλέον μέτρα που ορίστηκαν ως υπολογιζόμενα μέλη (Calculated Members) πάνω στον συγκεκριμένο κύβο, ορισμένα από τα οποία υπολογίστηκαν με βάση τα μέτρα του πίνακα 4-2. Τα Calculated Members υπολογίστηκαν με MDX Scripts τα οποία δίνονται στο Παράρτημα Α.5.

Πίνακας 4-1. Σύνοψη των διαστάσεων και ιεραρχιών του κύβου ΙΚΑ_ΟΑΕΕ_DW.cube

Διάσταση	Ιεραρχία	Περιγραφή
DIM_TIME	**ETOS** → **EXAMHNO**	Διάσταση Χρόνου
DIM_OIKONOM_ DRAST	Δεν έχει οριστεί Ιεραρχία	Διάσταση Οικονομικής Δραστηριότητας των ασφαλισμένων
DIM_YPOKATAST_ FOREAS	**ASFALISTIKO TAMEIO** → ASFAL FOREAS	Διάσταση Ασφαλιστικού Φορέα (ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ)
DIM_YPOKATAST_ FOREAS	**YPOKATASTHMA** → DESC YPOKATASHMATOS	Διάσταση Περιγραφής Υποκαταστημάτων Ασφαλιστικού Φορέα (ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ)
METANASTHS_OLAP	Δεν έχει οριστεί Ιεραρχία	Διάσταση με δημογραφικά στοιχεία του μετανάστη ασφαλισμένου
DIM_XWRA_KAT	**PROELEYSH METANASTH** → HPEIROS → ORGANISMOS → DESC_XVRA_KATAG → DESC_XWRA_KATAG_ENG	Διάσταση χώρας καταγωγής των μεταναστών ασφαλισμένων
DIM_OIKOGEN_ KATAST	**MARITAL STATUS** → DESC_OIKOG_KATAST	Διάσταση οικογενειακής κατάστασης
DIM_PER_KATOIKIAS	**TOPOTHESIA** → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής κατοικίας του ασφαλισμένου
DIM_PER_ERGASIAS	**TOPOTHESIA** → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής εργασίας του ασφαλισμένου
HLIKIAKES OMADES	**AGE_GROUPS** → HLIKIA (διακρίνεται στα διαστήματα) (17-24), (25-29), (30-32), (33-36), (37-40), (41-47) και (48-70)	Διάσταση ηλικιακών ομάδων. Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (DiscretizationMethod - EqualAreas)
PALAIOTHTA	**XRONOS_ASFALISHS** → ETH_ASFALISHS (διακρίνονται στα διαστήματα) (0-3), (4-7), (8-12), (13-16), (17-21), (22-26) και (27-46)	Διάσταση παλαιότητας, δηλαδή χρόνου ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα. Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (DiscretizationMethod - Clusters)
FYLO	**GENDER** → FYLO (Α,Γ)	Διάσταση φύλο ασφαλισμένου
METAVOLH ERGASIAS	Δεν έχει οριστεί Ιεραρχία	Διάσταση μεταβολή εργασίας

Πίνακας 4-2. Ομάδες μέτρων (measure groups) του κύβου IKA_OAEE_DW.cube

Measure Group	Μετρήσιμα Μεγέθη - Πεδία που δημιουργήθηκαν με τον Cube Wizard	Περιγραφή
PERIODIKH_KINHSH_METANASTH_OLAP	**PERIODIKH KINHSH METANASTH OLAP Count**	Πλήθος Μεταναστών του συγκεκριμένου πίνακα.
	APODOXES	Ποσά αποδοχών ασφαλισμένων (κενό πεδίο).
	EISFORES	Ποσά εισφορών ασφαλισμένων του ΟΑΕΕ.
	HLIKIA	Η ηλικία του ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται.
	ETH ASFALISHS	Τα έτη ασφάλισης κάθε ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται.
	SUM HLIKIA	Συνολική ηλικία ασφαλισμένων, ως βοηθητικό μέτρο υπολογισμού του μέσου όρου ηλικίας.
METANASTHS_OLAP	**ETOS GENNHSHS**	Έτος γέννησης ασφαλισμένων.
	ETOS APOGRAFHS	Έτος απογραφής ασφαλισμένων
	METANASTHS OLAP Count	Πλήθος Μεταναστών του συγκεκριμένου πίνακα.
PLITHOS_METANASTWN	**UNIQUE METANASTHS**	Εύρεση του διακριτού πλήθους των μεταναστών (με count distinct). Βοηθητικό μέτρο.

Πίνακας 4-3. Calculated Members πάνω στον κύβο ΙΚΑ_ΟΑΕΕ_DW.cube

Ορισμός μετρήσιμων μεγεθών σε όλο τον κύβο ως Calculated Members	Περιγραφή
ASFALISMENOI OAEE	Πλήθος ασφαλισμένων μεταναστών του ΟΑΕΕ. Βοηθητικό Μέτρο.
ASFALISMENOI IKA	Πλήθος ασφαλισμένων μεταναστών του ΙΚΑ_ΕΤΑΜ. Βοηθητικό Μέτρο.
MO HLIKIAS	Μέσος όρος ηλικίας του συνόλου των ασφαλισμένων.
MO EISFORWN	Μέσος όρος εισφορών ανά χρονική περίοδο των ασφαλισμένων του ΟΑΕΕ.
POSOSTO METANASTWN ANA OIKON DRAST	Ποσοστό ασφαλισμένων μεταναστών του ΙΚΑ-ΕΤΑΜ ανά οικονομική δραστηριότητα.
POSOSTO METANASTWN ANA ETHNIKOTHTA	Ποσοστό του συνόλου των ασφαλισμένων μεταναστών ανά εθνικότητα.

4.2 Παραδείγματα OLAP Ανάλυσης στον Κύβο

Στην προηγούμενη ενότητα ορίστηκε ο κύβος δεδομένων με τις διαστάσεις, τις ιεραρχίες και τα μετρήσιμα μεγέθη του. Στη συνέχεια, δίνονται με τη μορφή screenshot κάποια ενδεικτικά παραδείγματα αναλυτικής επεξεργασίας των δεδομένων του κύβου στο Analysis Services του Microsoft SQL Server 2005 με εφαρμογή διάφορων λειτουργιών OLAP για την εξαγωγή χρήσιμων συμπερασμάτων. Δηλαδή, σε κάθε ενδεικτικό παράδειγμα θα εκτελούνται κάποιες από τις επόμενες λειτουργίες και θα σχολιάζονται τα σχετικά αποτελέσματα.

Συσσώρευση (Roll-up): Πρόκειται για μια λειτουργία με την οποία εκτελείται ένα βήμα ανόδου στην ιεραρχία μιας διάστασης (π.χ. από ημέρα σε μήνα). Ο κύβος που προκύπτει από τη λειτουργία της συνάθροισης της πληροφορίας περιέχει πιο ομαδοποιημένα δεδομένα, με βάση τη διάσταση στην οποία έγινε η ομαδοποίηση. Η ανάβαση στην ιεραρχία μπορεί να συνεχιστεί με όμοιο τρόπο. Γενικά, η λειτουργία αυτή παρέχει συγκεντρωτικά αποτελέσματα τα οποία μπορούν να χρησιμοποιηθούν για την εξαγωγή στατιστικών στοιχείων για τα δεδομένα που αποθηκεύονται στην Αποθήκη.

Εμβάθυνση (Drill-down): Είναι η αντίστροφη πράξη του roll-up, όπου εκτελείται ένα βήμα καθόδου από ένα υψηλότερο επίπεδο της ιεραρχίας μιας διάστασης σε ένα χαμηλότερο. Με την εφαρμογή της συγκεκριμένης λειτουργίας, ο κύβος επιστρέφει αποτελέσματα με μεγάλο βαθμό λεπτομέρειας. Επίσης, η λειτουργία αυτή παρέχει τη δυνατότητα στον αναλυτή να διατρέξει ακόμη και ολόκληρη την ιεραρχία μιας κλάσης δεδομένων και να φτάσει στο χαμηλότερο επίπεδο λεπτομέρειας.

Τεμαχισμός (Slice): Πρόκειται για λειτουργία επιλογής δεδομένων σε μία συγκεκριμένη διάσταση. Ένα επίπεδο (slice) είναι ένα υποσύνολο ενός υπερκύβου σύμφωνα με μία περιοχή τιμών ή μια συγκεκριμένη τιμή ενός επιπέδου διάστασης (οριζόντιος τεμαχισμός). Ουσιαστικά, η συγκεκριμένη λειτουργία φιλτράρει τα αποτελέσματα που δίνει ο κύβος ως προς μια διάστασή του.

Κομμάτιασμα (Dice): Πρόκειται για λειτουργία επιλογής δεδομένων από δύο ή και περισσότερες διαστάσεις (κάθετος τεμαχισμός). Ουσιαστικά, η συγκεκριμένη λειτουργία φιλτράρει περισσότερες από μία διαστάσεις του κύβου για την εξαγωγή των επιθυμητών αποτελεσμάτων.

Περιστροφή (Pivot): Πρόκειται για λειτουργία αλλαγής της διάταξης των διαστάσεων ώστε να διευκολυνθεί η ανάλυση. Κατά την περιστροφή, δεν μεταβάλλονται ούτε μειώνονται τα

δεδομένα του υπερκύβου, απλά αλλάζει ο τρόπος παρουσίασής τους στην εφαρμογή ανάλυσης.

Παράδειγμα 1^{ο}: Μας ενδιαφέρει να μελετήσουμε ανά φορέα ασφάλισης, σε ποια γεωγραφικά διαμερίσματα του ελλαδικού χώρου διαμένουν και απασχολούνται οι περισσότεροι αλλοδαποί ασφαλισμένοι, άνδρες και γυναίκες, καθώς και από ποια ήπειρο προέρχονται.

			FYLO		
			Α	Γ	Grand Total
ASFAL FOREAS	DIAMERISMA	HPEIROS	METANASTHS OLAP Count	METANASTHS OLAP Count	METANASTHS OLAP Count
⊟ ΙΚΑ ΕΤΑΜ	⊟ ΣΤΕΡΕΑ ΕΛΛΑΔΑ	00	1	2	3
		AF	1311	276	1587
		AS	11035	2119	13154
		EU	17873	14482	32355
		NA	72	91	163
		OC	13	22	35
		SA	46	52	98
		Total	30351	17044	47395
	⊞ ΜΑΚΕΔΟΝΙΑ		3660	2510	6170
	⊞ ΠΕΛΟΠΟΝΝΗΣΟΣ		3240	2594	5834
	⊞ ΝΗΣΙΑ ΑΙΓΑΙΟΥ		1164	887	2051
	⊞ ΚΡΗΤΗ		825	688	1513
	⊞ ΘΕΣΣΑΛΙΑ		578	479	1057
	⊞ ΕΠΤΑΝΗΣΑ		267	302	569
	⊞ ΗΠΕΙΡΟΣ		292	267	559
	⊞ ΘΡΑΚΗ		162	174	336
	Total		40539	24945	65484
⊟ ΟΑΕΕ	⊟ ΣΤΕΡΕΑ ΕΛΛΑΔΑ	AF	1851	912	2763
		AS	5450	2253	7703
		EU	9812	7178	16990
		NA	1441	1144	2585
		OC	783	470	1253
		SA	209	205	414
		Total	19546	12162	31708
	⊞ ΜΑΚΕΔΟΝΙΑ		10877	7653	18530
	⊞ ΠΕΛΟΠΟΝΝΗΣΟΣ		2610	2078	4688
	⊞ ΝΗΣΙΑ ΑΙΓΑΙΟΥ		2120	1674	3794
	⊞ ΚΡΗΤΗ		1610	1323	2933
	⊞ ΘΕΣΣΑΛΙΑ		1600	1240	2840
	⊞ ΘΡΑΚΗ		1416	1245	2661
	⊞ ΗΠΕΙΡΟΣ		1191	735	1926
	⊞ ΕΠΤΑΝΗΣΑ		931	651	1582
	Total		41901	28761	70662
Grand Total			82440	53706	136146

Εικόνα 4-1. Εξαγόμενα αποτελέσματα 1ου παραδείγματος

Όπως φαίνεται στην παραπάνω εικόνα, οι περισσότεροι αλλοδαποί ασφαλισμένοι και των δύο φορέων ασφάλισης (ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ) διαμένουν και απασχολούνται στα γεωγραφικά διαμερίσματα της Στερεάς Ελλάδας και της Μακεδονίας και προέρχονται από τις ηπείρους της Ευρώπης και της Ασίας. Επομένως στο σύνολο των 136146 ασφαλισμένων (82440 ανδρών και 53706 γυναικών), οι 47395 απασχολούνται με εξαρτημένη εργασία (ασφαλισμένοι ΙΚΑ) και δραστηριοποιούνται στη Στερεά Ελλάδα και οι 31708 που απασχολούνται ως ελεύθεροι επαγγελματίες (ασφαλισμένοι του ΟΑΕΕ) δραστηριοποιούνται στο ίδιο γεωγραφικό διαμέρισμα. Ουσιαστικά, παρατηρούμε ότι ο ενεργός πληθυσμός των μεταναστών που δραστηριοποιούνται στη χώρα μας ακολουθεί την κατανομή του γενικού πληθυσμού, δηλαδή συσσωρεύεται στο γεωγραφικό διαμέρισμα της Στερεάς Ελλάδας, στο οποίο ανήκει ο πολυπληθέστερος νομός της χώρας, ο νομός Αττικής.

Παράδειγμα 2^{ο}: Μας ενδιαφέρει να μελετήσουμε από ποιες χώρες προέρχεται το μεγαλύτερο ποσοστό του συνόλου των ασφαλισμένων μεταναστών που εξετάζουμε, οι οποίοι βέβαια διαμένουν και απασχολούνται στα γεωγραφικά διαμερίσματα που συγκεντρώνουν τους περισσότερους μετανάστες.

		FYLO		
		A	Γ	Grand Total
DIAMERISMA	DESC XWRA KATAG	POSOSTO METANASTWN ANA ETHNIKOTHTA	POSOSTO METANASTWN ANA ETHNIKOTHTA	POSOSTO METANASTWN ANA ETHNIKOTHTA
ΜΑΚΕΔΟΝΙΑ	ΟΜΟΣΠ. ΔΗΜ. ΤΗΣ ΓΕΡΜΑΝΙΑΣ	45,93%	38,11%	42,74%
	ΑΛΒΑΝΙΑ	23,81%	16,23%	20,73%
	ΓΕΩΡΓΙΑ	13,68%	19,50%	16,05%
	ΡΩΣΙΑ	6,09%	10,32%	7,81%
	ΒΟΥΛΓΑΡΙΑ	3,70%	7,26%	5,15%
	ΑΥΣΤΡΑΛΙΑ	3,05%	3,33%	3,16%
	ΗΝΩΜΕΝΕΣ ΠΟΛ. ΤΗΣ ΑΜΕΡΙΚΗΣ	1,65%	2,00%	1,79%
	ΡΟΥΜΑΝΙΑ	1,41%	3,16%	2,13%
	ΠΑΚΙΣΤΑΝ	0,37%		0,22%
	ΙΝΔΙΑ	0,31%	0,07%	0,21%
	Total	100,00%	100,00%	100,00%
ΠΕΛΟΠΟΝΝΗΣΟΣ	ΑΛΒΑΝΙΑ	49,14%	42,56%	46,31%
	ΑΥΣΤΡΑΛΙΑ	10,18%	8,07%	9,27%
	ΟΜΟΣΠ. ΔΗΜ. ΤΗΣ ΓΕΡΜΑΝΙΑΣ	9,85%	8,10%	9,09%
	ΡΟΥΜΑΝΙΑ	8,37%	10,28%	9,19%
	ΒΟΥΛΓΑΡΙΑ	5,71%	11,77%	8,32%
	ΡΩΣΙΑ	4,79%	12,47%	8,09%
	ΗΝΩΜΕΝΕΣ ΠΟΛ. ΤΗΣ ΑΜΕΡΙΚΗΣ	4,57%	5,48%	4,96%
	ΙΝΔΙΑ	4,44%	0,09%	2,57%
	ΠΑΚΙΣΤΑΝ	2,07%	0,03%	1,19%
	ΓΕΩΡΓΙΑ	0,88%	1,17%	1,00%
	Total	100,00%	100,00%	100,00%
ΣΤΕΡΕΑ ΕΛΛΑΔΑ	ΑΛΒΑΝΙΑ	50,44%	47,92%	49,54%
	ΠΑΚΙΣΤΑΝ	16,54%	0,09%	10,64%
	ΙΝΔΙΑ	7,05%	0,31%	4,63%
	ΟΜΟΣΠ. ΔΗΜ. ΤΗΣ ΓΕΡΜΑΝΙΑΣ	5,96%	6,75%	6,25%
	ΡΟΥΜΑΝΙΑ	5,49%	9,39%	6,89%
	ΡΩΣΙΑ	5,48%	13,10%	8,21%
	ΒΟΥΛΓΑΡΙΑ	2,93%	11,57%	6,03%
	ΗΝΩΜΕΝΕΣ ΠΟΛ. ΤΗΣ ΑΜΕΡΙΚΗΣ	2,55%	3,88%	3,03%
	ΑΥΣΤΡΑΛΙΑ	2,26%	2,50%	2,35%
	ΓΕΩΡΓΙΑ	1,30%	4,49%	2,44%
	Total	100,00%	100,00%	100,00%
Grand Total		100,00%	100,00%	100,00%

Εικόνα 4-2. Εξαγόμενα αποτελέσματα 2ου παραδείγματος

Όπως φαίνεται στην παραπάνω εικόνα, στο γεωγραφικό διαμέρισμα της Μακεδονίας το μεγαλύτερο ποσοστό μεταναστών (ανδρών και γυναικών) προέρχεται από την Ομόσπονδη Δημοκρατία της Γερμανίας με ποσοστό 42,74% επί του συνόλου των ασφαλισμένων στο συγκεκριμένο γεωγραφικό διαμέρισμα, και έπονται οι Αλβανοί μετανάστες με αρκετά σημαντικό ποσοστό 20,73% επί του συνόλου. Στο γεωγραφικό διαμέρισμα της Πελοποννήσου οι περισσότεροι μετανάστες προέρχονται από την Αλβανία με ποσοστό 46,31% επί του συνόλου. Τέλος, στο γεωγραφικό διαμέρισμα της Στερεάς Ελλάδας το μεγαλύτερο ποσοστό των μεταναστών 49,54% επί του συνόλου προέρχεται και πάλι από την Αλβανία, και έπονται οι μετανάστες από το Πακιστάν με ποσοστό 10,64% επί του συνόλου.

Παράδειγμα 3^{ο}: Μας ενδιαφέρει να μελετήσουμε τους μετανάστες που προέρχονται από τις χώρες της Αλβανίας και της Ομόσπονδης Δημοκρατίας της Γερμανίας (μεγαλύτερο ποσοστό ασφαλισμένων μεταναστών επί του συνόλου), ως προς χρόνο ασφάλισής τους στους ασφαλιστικούς φορείς και το μέσο όρο ηλικίας τους, συγκρίνοντας το πρώτο εξάμηνο του έτους 2007 με πρώτο εξάμηνο του έτους 2008.

EXAMHNO								
1								
			ETOS					
			2007		2008		Grand Total	
DESC XWRA KATAG	FYLO	ETH ASFALISHS	METANASTHS OLAP Count	MO HLIKIAS	METANASTHS OLAP Count	MO HLIKIAS	METANASTHS OLAP Count	MO HLIKIAS
⊟ ΑΛΒΑΝΙΑ	⊟ Α	0 - 3	**5988**	**28,9**	**6539**	**30,2**	6698	29,6
		4 - 7	**7954**	**32,1**	**7897**	**33,0**	7972	32,6
		8 - 12	**6154**	**38,3**	**6154**	**39,3**	6162	38,8
		13 - 16	**3027**	**41,3**	**3027**	**42,3**	3027	41,8
		17 - 21	**112**	**42,3**	**112**	**43,3**	112	42,8
		22 - 26	**8**	**55,0**	**7**	**54,3**	8	54,7
		27 - 46	**10**	**57,9**	**11**	**59,1**	11	58,5
		Total	23253	34,2	23747	35,1	23990	34,7
	⊟ Γ	0 - 3	**3364**	**28,9**	**3662**	**30,2**	3779	29,5
		4 - 7	**4049**	**31,5**	**3998**	**32,4**	4053	32,0
		8 - 12	**3236**	**38,4**	**3235**	**39,4**	3240	38,9
		13 - 16	**1824**	**41,5**	**1824**	**42,5**	1824	42,0
		17 - 21	**69**	**42,2**	**69**	**43,2**	69	42,7
		22 - 26	**2**	**53,5**	**2**	**54,5**	2	54,0
		27 - 46	**3**	**60,3**	**3**	**61,3**	3	60,8
		Total	12547	34,1	12793	35,1	12970	34,6
	Total		35800	34,2	36540	35,1	36960	34,6
⊟ ΟΜΟΣΠ. ΔΗΜ. ΤΗΣ ΓΕΡΜΑΝΙΑΣ	⊟ Α	0 - 3	**2243**	**34,8**	**2732**	**35,4**	2979	35,2
		4 - 7	**2660**	**35,8**	**2538**	**36,7**	2694	36,2
		8 - 12	**2115**	**36,9**	**2077**	**37,9**	2173	37,4
		13 - 16	**1024**	**39,0**	**1030**	**40,0**	1062	39,5
		17 - 21	**648**	**41,2**	**647**	**42,0**	674	41,6
		22 - 26	**128**	**43,9**	**141**	**44,8**	141	44,4
		27 - 46	**15**	**54,3**	**14**	**55,1**	15	54,7
		Total	8833	36,7	9179	37,5	9738	37,1
	⊟ Γ	0 - 3	**1641**	**35,2**	**1947**	**35,9**	2179	35,5
		4 - 7	**1666**	**36,5**	**1533**	**37,5**	1683	37,0
		8 - 12	**938**	**38,0**	**905**	**39,0**	956	38,5
		13 - 16	**439**	**39,9**	**438**	**40,8**	453	40,4
		17 - 21	**295**	**41,5**	**294**	**42,2**	310	41,8
		22 - 26	**86**	**45,9**	**90**	**46,5**	93	46,2
		27 - 46	**15**	**58,1**	**13**	**59,5**	15	58,8
		Total	5080	37,1	5220	37,9	5689	37,5
	Total		13913	36,9	14399	37,6	15427	37,3
Grand Total			49713	34,9	50939	35,8	52387	35,4

Εικόνα 4-3. Εξαγόμενα αποτελέσματα 3ου παραδείγματος

Όπως φαίνεται στην παραπάνω εικόνα, συγκρίνοντας το πλήθος των ασφαλισμένων και το μέσο όρο ηλικίας τους που αντιστοιχούν στα διαστήματα των ετών ασφάλισης στους ασφαλιστικούς φορείς, παρατηρούμε ότι οι Αλβανοί μετανάστες ξεκινούν την ενεργή επαγγελματική τους δραστηριότητα σε μικρότερη ηλικία από τους Γερμανούς μετανάστες, με διαφορά πέντε ετών περίπου. Επίσης, ο μέσος όρος ηλικίας των γυναικών που απασχολούνται δεν διαφέρει σημαντικά με το μέσο όρο ηλικίας των ανδρών. Ωστόσο, σε κάθε περίπτωση για τις συγκεκριμένες κατηγορίες των ασφαλισμένων που μελετάμε παρατηρούμε ότι η παραγωγική ηλικία ξεκινά περίπου στα 25 έτη για τους Αλβανούς ασφαλισμένους και στα 30 έτη για τους Γερμανούς.

Παράδειγμα 4^{ο}: Σε συνέχεια του Παραδείγματος 3, για την ίδια χρονική περίοδο, μας ενδιαφέρει να μελετήσουμε την έναρξη της παραγωγικής ηλικίας μόνο των Αλβανών μεταναστών στους δύο ασφαλιστικούς φορείς.

EXAMHNO						
1						
				ETOS		
				2007	2008	Grand Total
ASFAL FOREAS	DESC XWRA KATAG	FYLO	ETH ASFALISHS	MO HLIKIAS	MO HLIKIAS	MO HLIKIAS
ΙΚΑ ΕΤΑΜ	ΑΛΒΑΝΙΑ	Α	0 - 3	**26,4**	**27,4**	26,9
			4 - 7	**29,7**	**30,7**	30,2
			8 - 12	**37,9**	**38,9**	38,4
			13 - 16	**41,2**	**42,2**	41,7
			17 - 21	**41,8**	**42,8**	42,3
			Total	33,2	34,2	33,7
		Γ	0 - 3	**26,2**	**27,2**	26,7
			4 - 7	**29,7**	**30,7**	30,2
			8 - 12	**38,3**	**39,3**	38,8
			13 - 16	**41,5**	**42,5**	42,0
			17 - 21	**41,9**	**42,9**	42,4
			Total	33,5	34,5	34,0
		Total		33,3	34,3	33,8
	Total			33,3	34,3	33,8
ΟΑΕΕ	ΑΛΒΑΝΙΑ	Α	0 - 3	**35,4**	**35,8**	35,6
			4 - 7	**39,2**	**40,1**	39,7
			8 - 12	**42,5**	**43,5**	43,0
			13 - 16	**47,3**	**48,3**	47,8
			17 - 21	**48,4**	**49,4**	48,9
			22 - 26	**55,0**	**54,3**	54,7
			27 - 46	**57,9**	**59,1**	58,5
			Total	38,4	38,7	38,6
		Γ	0 - 3	**35,6**	**35,9**	35,8
			4 - 7	**38,8**	**39,8**	39,3
			8 - 12	**40,8**	**41,9**	41,4
			13 - 16	**44,7**	**45,7**	45,2
			17 - 21	**51,0**	**52,0**	51,5
			22 - 26	**53,5**	**54,5**	54,0
			27 - 46	**60,3**	**61,3**	60,8
			Total	37,6	37,9	37,8
		Total		38,2	38,5	38,3
	Total			38,2	38,5	38,3
Grand Total				34,2	35,1	34,6

Εικόνα 4-4. Εξαγόμενα αποτελέσματα 4ου παραδείγματος

Όπως φαίνεται στα εξαγόμενα αποτελέσματα, τόσο οι άνδρες όσο και οι γυναίκες που προέρχονται από την Αλβανία και είναι ασφαλισμένοι στο ΙΚΑ ξεκινούν να εργάζονται σε μικρότερη ηλικία από τους ασφαλισμένους του ΟΑΕΕ και η διαφορά κυμαίνεται περίπου μεταξύ των εννέα και δέκα ετών. Επομένως, η άσκηση εξαρτημένης εργασίας ξεκινά για το συγκεκριμένο πληθυσμό σε μικρότερη ηλικία (25 ετών περίπου) ενώ άτομα μεγαλύτερης ηλικίας (35 ετών περίπου) αποφασίζουν να απασχοληθούν ως ελεύθεροι επαγγελματίες, γεγονός που δείχνει ότι η ηλικία επηρεάζει σε μεγάλο βαθμό την επιλογή απασχόλησης.

Παράδειγμα 5ο: Μας ενδιαφέρει να μελετήσουμε την εξέλιξη του μέσου όρου των εξαμηνιαίων εισφορών των ασφαλισμένων του ΟΑΕΕ ανά ηλικιακή ομάδα και για τη χρονική περίοδο τριών εξαμήνων, Ιούνιος 2007 έως Ιούνιος 2009.

	ΕΤΟΣ ▾ ΕΞΑΜΗΝΟ									
	⊟2007				⊟2008				⊟2009	
	1		Total		1		Total		1	
ΗΛΙΚΙΑ ▾	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΕΙΣΦΟΡΩΝ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΕΙΣΦΟΡΩΝ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΕΙΣΦΟΡΩΝ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΕΙΣΦΟΡΩΝ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΕΙΣΦΟΡΩΝ
17 - 24	961	813,41	961	813,41	1366	953,07	1366	953,07	1640	1.100,69
25 - 29	4155	922,97	4155	922,97	4939	1.049,49	4939	1.049,49	5482	1.163,57
30 - 32	5022	1.048,94	5022	1.048,94	5548	1.167,25	5548	1.167,25	5856	1.266,00
33 - 36	10408	1.201,12	10408	1.201,12	11010	1.314,18	11010	1.314,18	11369	1.397,49
37 - 40	10096	1.322,82	10096	1.322,82	10582	1.438,22	10582	1.438,22	10949	1.507,86
41 - 47	12884	1.392,69	12884	1.392,69	13348	1.514,07	13348	1.514,07	13677	1.579,56
48 - 70	11911	1.477,54	11911	1.477,54	12028	1.603,31	12028	1.603,31	11927	1.670,36
Grand Total	55437	1.285,84	55437	1.285,84	58821	1.396,51	58821	1.396,51	60900	1.469,97

Εικόνα 4-5. Εξαγόμενα αποτελέσματα 5ου παραδείγματος

Όπως ήταν αναμενόμενο, στις μικρές ηλικίες οι καταβληθείσες εξαμηνιαίες εισφορές είναι μικρότερες συγκριτικά με αυτές στις μεγαλύτερες ηλικιακές κλάσεις. Επίσης, οι εισφορές ακολουθούν αυξητική πορεία ως προς το χρόνο. Τέλος, παρατηρούμε ότι σε κάθε εξάμηνο το σύνολο των ασφαλισμένων μεταβάλλεται χωρίς ποτέ να φτάνει τον πραγματικό αριθμό ασφαλισμένων του ΟΑΕΕ που εξετάζουμε, δηλαδή τους 70662 ασφαλισμένους. Αυτό συμβαίνει, επειδή τα διαθέσιμα οικονομικά στοιχεία για την τριετία 2007-2009 αντιπροσωπεύουν τις εξαμηνιαίες εισφορές των τακτικών ασφαλισμένων, δηλαδή αυτών που δε χρωστούν κάποια δόση.

Παράδειγμα 6ο: Μας ενδιαφέρει να μελετήσουμε τους ασφαλισμένους του ΟΑΕΕ ως προς την οικογενειακή τους κατάσταση και το μέσο όρο ηλικίας τους για την τριετία 2007 έως 2009, κατά την οποία υπάρχουν σχετικά διαθέσιμα στοιχεία.

	ΕΤΟΣ ▾							
	2007		2008		2009		Grand Total	
DESC ΟΙΚΟΓ ΚΑΤΑΣΤ ▾	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΗΛΙΚΙΑΣ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΗΛΙΚΙΑΣ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΗΛΙΚΙΑΣ	ΑΣΦΑΛΙΣΜΕΝΟΙ ΟΑΕΕ	ΜΟ ΗΛΙΚΙΑΣ
ΕΓΓΑΜΟΣ	32463	42,1	33471	42,6	33312	43,3	37727	42,7
ΑΓΑΜΟΣ	24703	35,6	26616	36,2	26969	36,9	30443	36,3
ΔΙΑΖΕΥΜΕΝΟΣ	1754	44,0	1808	44,6	1773	45,2	2072	44,6
ΧΗΡΟΣ	372	48,5	373	48,9	358	49,7	420	49,0
Grand Total	59292	39,5	62268	40,0	62412	40,7	70662	40,1

Εικόνα 4-6. Εξαγόμενα αποτελέσματα 6ου παραδείγματος

Όπως παρατηρούμε, κατά την τριετία 2007 έως 2009 οι περισσότεροι ασφαλισμένοι του ΟΑΕΕ, 37727 ασφαλισμένοι επί του συνόλου των 70662, είναι έγγαμοι με μέσο όρο ηλικίας τα 42,7 έτη.

Παράδειγμα 7^{ο}: Μας ενδιαφέρει να μελετήσουμε τους ασφαλισμένους του ΙΚΑ (άνδρες και γυναίκες) ως προς τις κυριότερες κατηγορίες οικονομικής δραστηριότητας στις οποίες απασχολούνται για τα έτη 2007, 2008 καθώς και τις αντίστοιχες ηλικιακές ομάδες στις οποίες ανήκει το μεγαλύτερο ποσοστό αυτών.

			ΕΤΟΣ	
			2007	2008
FYLO	DESC OIKONOM DRAST	HLIKIA	POSOSTO METANASTWN	POSOSTO ME
Α	ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ	48 - 70	53,83%	50,57%
		25 - 29	43,40%	40,86%
		17 - 24	36,35%	33,87%
		Total	42,53%	39,86%
	ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ	17 - 24	25,79%	23,20%
		25 - 29	22,92%	19,29%
		48 - 70	18,88%	15,07%
		Total	23,31%	20,05%
	ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ - ΕΠΙΣΚΕΥΗ ΑΥΤΟΚΙΝΗΤΩΝ ΟΧΗΜΑΤΩΝ, ΜΟΤΟΣΥΚΛΕΤΩΝ ΚΑΙ ΕΙΔΩΝ ΠΡΟΣΩΠΙΚΗΣ ΚΑΙ ΟΙΚΙΑΚΗΣ ΧΡΗΣΗΣ	25 - 29	29,12%	28,40%
		17 - 24	28,68%	27,80%
		48 - 70	20,26%	19,61%
		Total	27,43%	26,68%
	ΚΑΤΑΣΚΕΥΕΣ	17 - 24	16,73%	14,73%
		25 - 29	12,43%	10,50%
		48 - 70	12,12%	10,18%
		Total	14,00%	12,05%
	ΙΔΙΩΤΙΚΑ ΝΟΙΚΟΚΥΡΙΑ ΠΟΥ ΑΠΑΣΧΟΛΟΥΝ ΟΙΚΙΑΚΟ ΠΡΟΣΩΠΙΚΟ ΚΑΙ ΜΗ ΔΙΑΦΟΡΟΠΟΙΗΜΕΝΕΣ ΠΑΡΑΓΩΓΙΚΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ ΝΟΙΚΟΚΥΡΙΩΝ ΓΙΑ ΙΔΙΑ ΧΡΗΣΗ	48 - 70	5,30%	4,57%
		25 - 29	1,05%	0,96%
		17 - 24	0,42%	0,40%
		Total	1,54%	1,36%
	Total		100,00%	100,00%
Γ	ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ	25 - 29	49,60%	47,53%
		17 - 24	47,61%	44,52%
		48 - 70	29,05%	34,93%
		Total	42,49%	42,57%
	ΙΔΙΩΤΙΚΑ ΝΟΙΚΟΚΥΡΙΑ ΠΟΥ ΑΠΑΣΧΟΛΟΥΝ ΟΙΚΙΑΚΟ ΠΡΟΣΩΠΙΚΟ ΚΑΙ ΜΗ ΔΙΑΦΟΡΟΠΟΙΗΜΕΝΕΣ ΠΑΡΑΓΩΓΙΚΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ ΝΟΙΚΟΚΥΡΙΩΝ ΓΙΑ ΙΔΙΑ ΧΡΗΣΗ	48 - 70	59,21%	47,01%
		25 - 29	9,65%	8,41%
		17 - 24	5,24%	4,56%
		Total	23,75%	19,31%
	ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ - ΕΠΙΣΚΕΥΗ ΑΥΤΟΚΙΝΗΤΩΝ ΟΧΗΜΑΤΩΝ, ΜΟΤΟΣΥΚΛΕΤΩΝ ΚΑΙ ΕΙΔΩΝ ΠΡΟΣΩΠΙΚΗΣ ΚΑΙ ΟΙΚΙΑΚΗΣ ΧΡΗΣΗΣ	17 - 24	35,84%	35,09%
		25 - 29	26,17%	25,48%
		48 - 70	8,20%	8,04%
		Total	23,73%	23,15%
	ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ	25 - 29	19,65%	17,66%
		17 - 24	15,90%	14,79%
		48 - 70	9,60%	9,14%
		Total	15,25%	14,03%
	ΚΑΤΑΣΚΕΥΕΣ	17 - 24	1,48%	1,05%
		25 - 29	1,33%	0,92%
		48 - 70	1,24%	0,87%
		Total	1,35%	0,95%
	Total		100,00%	100,00%
Grand Total			100,00%	100,00%

Εικόνα 4-7. Εξαγόμενα αποτελέσματα 7ου παραδείγματος

Όπως παρατηρούμε, οι κυριότερες κατηγορίες οικονομικής δραστηριότητας των ασφαλισμένων του ΙΚΑ για τα έτη 2007 και 2008 είναι οι ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ, ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ, ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ, ΚΑΤΑΣΚΕΥΕΣ και ΙΔΙΩΤΙΚΑ ΝΟΙΚΟΚΥΡΙΑ ΠΟΥ ΑΠΑΣΧΟΛΟΥΝ ΟΙΚΙΑΚΟ ΠΡΟΣΩΠΙΚΟ. Όσον αφορά στους άνδρες ασφαλισμένους, το μεγαλύτερο ποσοστό (περίπου 53%) που ανήκει και στην ηλικιακή ομάδα 48-70 απασχολείται στις ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ, και έπονται με ποσοστό περίπου 29% οι άνδρες ηλικίας μεταξύ 25-29 ετών που απασχολούνται στο ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ. Όσον αφορά στις γυναίκες ασφαλισμένες, και πάλι το μεγαλύτερο ποσοστό (περίπου 58%) που ανήκει και στην ηλικιακή ομάδα 48-70 απασχολείται σε ΙΔΙΩΤΙΚΑ ΝΟΙΚΟΚΥΡΙΑ ΠΟΥ ΑΠΑΣΧΟΛΟΥΝ ΟΙΚΙΑΚΟ ΠΡΟΣΩΠΙΚΟ, έπονται με ποσοστό περίπου 55% οι ΓΥΝΑΙΚΕΣ ηλικίας μεταξύ 25-29 ετών που απασχολούνται σε ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ, και τέλος 35% περίπου των γυναικών ηλικίας μεταξύ 17-24 απασχολούνται στο ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ.

Παράδειγμα 8^{ο}: Μας ενδιαφέρει να μελετήσουμε τους ασφαλισμένους του ΙΚΑ (άνδρες και γυναίκες) που διαμένουν και δραστηριοποιούνται στα γεωγραφικά διαμερίσματα της Μακεδονίας και της Στερεάς Ελλάδας, τα οποία συγκεντρώνουν τους περισσότερους ασφαλισμένους μετανάστες της χώρας, ως προς τις μεταβολές της οικονομικής τους δραστηριότητας. Αξίζει να σημειωθεί, ότι για τη χρονική περίοδο Ιούνιος 2004 έως Ιούνιος 2008 παράξαμε στοιχεία οικονομικής δραστηριότητας για το σύνολο των 65500 περίπου ασφαλισμένων, σύμφωνα με τα επίσημα δημοσιευμένα εξαμηνιαία στατιστικά δελτία του ΙΚΑ-ΕΤΑΜ.

				ETOS					
				2004	2005	2006	2007	2008	Grand Total
DIAMERISMA	FYLO	METAVOLES OIK DRAST	ETH ASFALISHS	METANASTHS OLAP Count	METANASTHS OLAP Count	METANASTHS OLAP Count	METANASTHS OLAP Count	METANASTHS OLAP Count	METANASTHS OLAP C
ΜΑΚΕΔΟΝΙΑ	Α	STABLE IKA	4 - 7	**796**	**796**	**796**	**796**	**796**	796
			0 - 3	**566**	**566**	**566**	**566**	**566**	566
			8 - 12	**512**	**512**	**512**	**512**	**512**	512
			Total	1874	1874	1874	1874	1874	1874
		IKA CHANGED	4 - 7	**589**	**589**	**589**	**589**	**589**	589
			0 - 3	**424**	**424**	**424**	**424**	**424**	424
			8 - 12	**387**	**387**	**387**	**387**	**387**	387
			Total	1400	1400	1400	1400	1400	1400
		Total		3274	3274	3274	3274	3274	3274
	Γ	IKA CHANGED	4 - 7	**491**	**491**	**491**	**491**	**491**	491
			8 - 12	**370**	**370**	**370**	**370**	**370**	370
			0 - 3	**323**	**323**	**323**	**323**	**323**	323
			Total	1184	1184	1184	1184	1184	1184
		STABLE IKA	4 - 7	**417**	**417**	**417**	**417**	**417**	417
			8 - 12	**318**	**318**	**318**	**318**	**318**	318
			0 - 3	**273**	**273**	**273**	**273**	**273**	273
			Total	1008	1008	1008	1008	1008	1008
		Total		2192	2192	2192	2192	2192	2192
	Total			5466	5466	5466	5466	5466	5466
ΣΤΕΡΕΑ ΕΛΛΑΔΑ	Α	STABLE IKA	4 - 7	**6314**	**6314**	**6314**	**6314**	**6314**	6314
			8 - 12	**5130**	**5130**	**5130**	**5130**	**5130**	5130
			0 - 3	**3619**	**3619**	**3619**	**3619**	**3619**	3619
			Total	15063	15063	15063	15063	15063	15063
		IKA CHANGED	4 - 7	**4565**	**4565**	**4565**	**4565**	**4565**	4565
			8 - 12	**3977**	**3977**	**3977**	**3977**	**3977**	3977
			0 - 3	**2729**	**2729**	**2729**	**2729**	**2729**	2729
			Total	11271	11271	11271	11271	11271	11271
		Total		26334	26334	26334	26334	26334	26334
	Γ	IKA CHANGED	4 - 7	**3186**	**3186**	**3186**	**3186**	**3186**	3186
			8 - 12	**2694**	**2694**	**2694**	**2694**	**2694**	2694
			0 - 3	**1907**	**1907**	**1907**	**1907**	**1907**	1907
			Total	7787	7787	7787	7787	7787	7787
		STABLE IKA	4 - 7	**2806**	**2806**	**2806**	**2806**	**2806**	2806
			8 - 12	**2257**	**2257**	**2257**	**2257**	**2257**	2257
			0 - 3	**1712**	**1712**	**1712**	**1712**	**1712**	1712
			Total	6775	6775	6775	6775	6775	6775
		Total		14562	14562	14562	14562	14562	14562
	Total			40896	40896	40896	40896	40896	40896
Grand Total				46362	46362	46362	46362	46362	46362

Εικόνα 4-8. Εξαγόμενα αποτελέσματα 8ου παραδείγματος

Όπως παρατηρούμε, και στα δύο γεωγραφικά διαμερίσματα οι περισσότεροι άνδρες διατηρούν την οικονομική τους δραστηριότητα ενώ στις γυναίκες η οικονομική δραστηριότητα μεταβάλλεται. Επίσης, παρόλο που οι περισσότερες γυναίκες είναι ασφαλισμένες για αρκετά μεγάλο χρονικό διάστημα στο συγκεκριμένο φορέα από 4 έως 12 έτη, παρατηρούμε ότι κυρίως αυτές αντιμετωπίζουν μεταβολές στην οικονομική τους δραστηριότητα, γεγονός που καταδεικνύει την αβεβαιότητα στην απασχόληση των γυναικών στη χώρα μας.

5. Μοντέλα Εξόρυξης Γνώσης (Data mining Models)

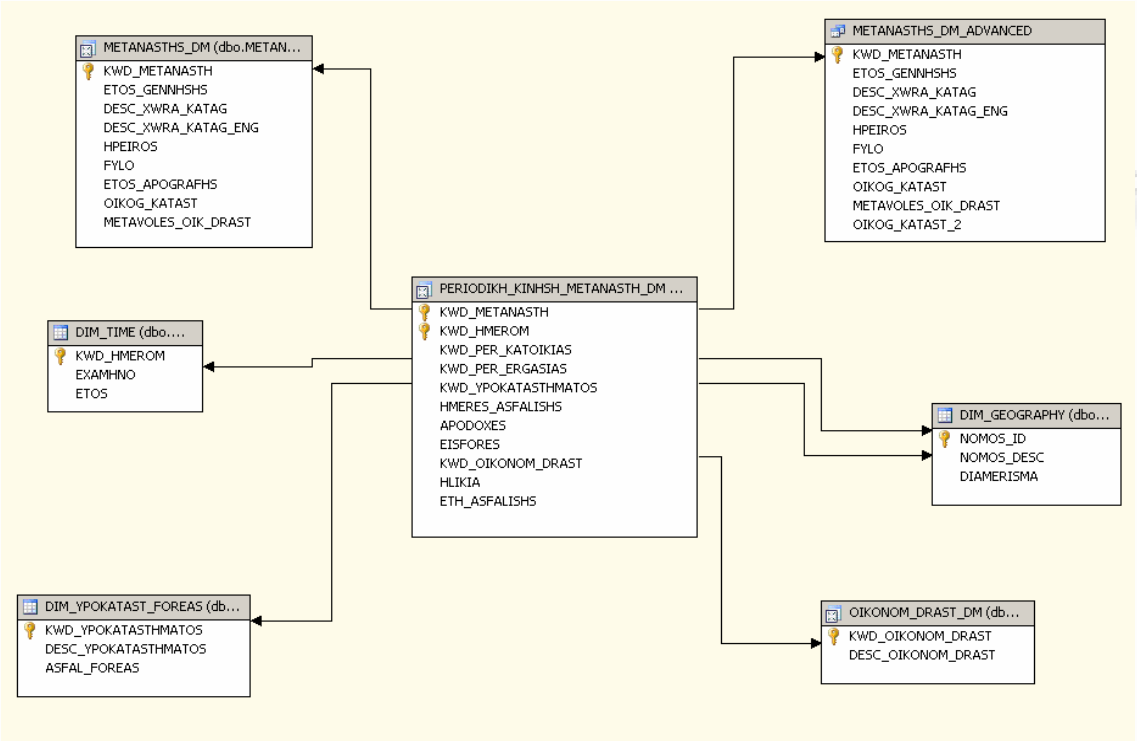
Στη συγκεκριμένη ενότητα παρουσιάζονται τεχνικές εξόρυξης γνώσης που εφαρμόστηκαν στα δεδομένα της Αποθήκης Μεταναστευτικών Δεδομένων. Ουσιαστικά, χρησιμοποιήθηκαν κάποιοι από του διαθέσιμους αλγόριθμους εξαγωγής προτύπων του εργαλείου Analysis Services του Microsoft SQL Server 2005. Ωστόσο, και πάλι λόγω της ιδιαιτερότητας που παρουσιάζουν τα διαθέσιμα δεδομένα των δύο ασφαλιστικών οργανισμών (βλ. Κεφάλαιο 4), σε κάθε επίπεδο ανάλυσης επιλέχθηκε μια πιο ευέλικτη μορφή (data source) της αρχικής Αποθήκης Δεδομένων με κατάλληλη προπαρασκευή των δεδομένων. Στη συνέχεια, παραθέτονται ενδεικτικά παραδείγματα αναλυτικής επεξεργασίας του μεγάλου όγκου της συγκεντρωμένης ιστορικής πληροφορίας στην Αποθήκη Δεδομένων καθώς και ενδεικτικά μοντέλα εφαρμογής τεχνικών Κατηγοριοποίησης (Classification), Συσταδοποίησης (Clustering) και Κανόνων Συσχετίσεων (Association Rules) σε υποθετικά σενάρια.

5.1 Προπαρασκευή της Αποθήκης Δεδομένων για τη Δημιουργία Μοντέλων Εξόρυξης Γνώσης

Η εφαρμογή τεχνικών εξόρυξης γνώσης σε ένα πολυδιάστατο μοντέλο δεδομένων – κύβο παρέχει δυνατότητες για χρησιμοποίηση των ιεραρχιών του κύβου και των λειτουργιών OLAP (π.χ. χρησιμοποίηση της λειτουργίας Slice – τεμαχισμός ή Dice – κομμάτιασμα), όμως ο κύβος αποδεικνύεται λιγότερο ευέλικτος από ένα κλασσικό σχεσιακό μοντέλο δεδομένων. Ωστόσο, για τις ανάγκες της παρούσας εργασίας τα μοντέλα εξόρυξης γνώσης που δημιουργήθηκαν με το εργαλείο Analysis Services του Microsoft SQL Server 2005 βασίστηκαν πάνω σε κατάλληλα δομημένους κύβους.

Επομένως, στα πλαίσια σχεδιασμού κατάλληλων κύβων με διαστάσεις, ιεραρχίες και μέτρα για την εφαρμογή διαδικασιών Κατηγοριοποίησης (Classification), Συσταδοποίησης (Clustering) και των Κανόνων Συσχετίσεων (Association Rules), επιλέχθηκε μια πιο ευέλικτη μορφή πηγής δεδομένων (data source) από αυτή της αρχικής Αποθήκης Μεταναστευτικών Δεδομένων της Παραγράφου 3.3.2. Η σχεσιακή αναπαράσταση της νέας Αποθήκης δίνεται στο παρακάτω Σχήμα 5-1. Στα Σχήματα 5-2 και 5-3 που ακολουθούν δίνονται οι κύβοι που δημιουργήθηκαν πάνω στο data source της Αποθήκης τους Σχήματος 5-1.

Η δομή της Αποθήκης Δεδομένων όπως απεικονίζεται στο Σχήμα 5-1 διαφέρει σε αρκετά σημεία σε σχέση με τη δομή που είχε επιλεχθεί για το σχεσιακό σχήμα της Παραγράφου 3.3.2., και προήλθε κυρίως από την ανάγκη για μεγαλύτερη ευελιξία των κύβων που θα δημιουργηθούν πάνω σε αυτή και στη συνέχεια θα χρησιμοποιηθούν στις διαδικασίες εξόρυξης γνώσης. Πρέπει επίσης να σημειωθεί ότι το νέο data source προέκυψε και από παρεμβάσεις που έγιναν στους πίνακες και στα πεδία τους ώστε συμπληρωθούν οι ελλιπείς τιμές, καθώς οι αλγόριθμοι εξόρυξης γνώσης του Microsoft SQL Server 2005 δεν μπορούν να εφαρμοστούν όταν υπάρχουν ελλιπείς τιμές σε πεδία πινάκων διαστάσεων, ενώ δεν επηρεάζονται από ελλιπείς τιμές στα μετρήσιμα μεγέθη των κύβων, αν αυτά δεν χρησιμοποιηθούν στα διάφορα μοντέλα. Το νέο σχεσιακό σχήμα έχει τη δομή Σχήματος Αστέρα (Star Schema), με κεντρικό πίνακα γεγονότων τον PERIODIKH_KINHSH_METANASTH_DM και κάποιους από-κανονικοποιημένους πίνακες διαστάσεων, μερικοί εκ των οποίων προέκυψαν από συγχώνευση άλλων πινάκων.



Σχήμα 5-1. Η Αποθήκη Δεδομένων IKA_OAEE_DW.ds, που θα χρησιμοποιηθεί ως data source στους κύβους IKA_OAEE_DW_SIMPLE.cube και IKA_OAEE_DW_ADVANCED.cube

Ο πίνακας γεγονότων PERIODIKH_KINHSH_METANASTH_DM αποτελεί μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα FACT_PERIODIKH_KINHSH της αρχικής Αποθήκης, ενώ παράλληλα διαθέτει δύο επιπλέον πεδία – μέτρα, τα [HLIKIA] και [ETH_ASFALISHS]. Στα πεδία αυτά, τόσο η ηλικία όσο και τα έτη ασφάλισης, δηλαδή ο χρόνος ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα, υπολογίζονται για τις χρονικές περιόδους που καλύπτει η Αποθήκη Δεδομένων, δηλαδή για την περίοδο Ιούνιος 2004 έως Ιούνιος 2008 για τους ασφαλισμένους του ΙΚΑ_ΕΤΑΜ και για την περίοδο Ιούνιος 2007 έως Δεκέμβριος 2009 για τους ασφαλισμένους του ΟΑΕΕ, όπως δηλαδή και στον πίνακα PERIODIKH_KINHSH_METANASTH_OLAP. Μια επιπλέον διαφορά του συγκεκριμένου πίνακα είναι εισαγωγή της τιμής “0” αντί της NULL στον [KWD_OIKONOM_DRAST] για τους ασφαλισμένους του ΟΑΕΕ, για τους οποίους δεν υπάρχουν διαθέσιμα στοιχεία οικονομικής δραστηριότητας. Παράλληλα, και ο πίνακας διάστασης OIKONOM_DRAST_DM αποτελεί μια όψη στην οποία χρησιμοποιήθηκε η γενική σταθερά “ΑΓΝΩΣΤΟ” ως περιγραφή του κωδικού οικονομικής δραστηριότητας “0”, με σκοπό τη διαχείριση των ελλιπών τιμών οικονομικής δραστηριότητας για τους ασφαλισμένους του ΟΑΕΕ. Τα SQL Script δημιουργίας των views δίνονται στο Παράρτημα Α.4.

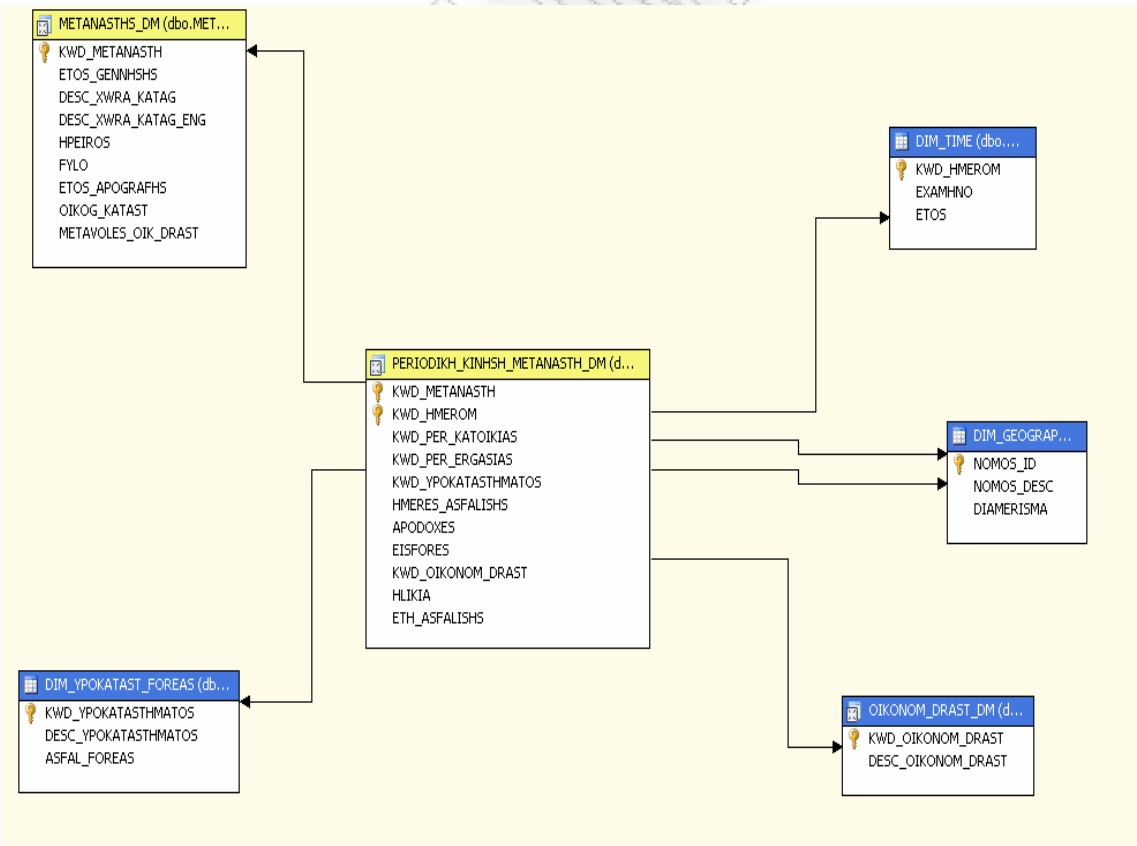
Όσον αφορά στους υπόλοιπους από-κανονικοποιημένους πίνακες διαστάσεων του data source του Σχήματος 5-1, παρατηρούμε ότι διατηρούνται οι διαστάσεις DIM_TIME, DIM_GEOGRAPHY και DIM_YPOKATAST_FOREAS της αρχικής Αποθήκης Δεδομένων ενώ υπάρχουν δύο επιπλέον πίνακες, οι METANASTHS_DM και METANASTHS_DM_ADVANCED στους οποίους έχουν συγχωνευτεί πεδία των αρχικών πινάκων διαστάσεων DIM_XWRA_KAT και DIM_OIKOGEN_KATAST. Η συγχώνευση αυτή προήλθε από την ανάγκη δημιουργίας πιο ευέλικτων κύβων για την εφαρμογή διαδικασιών εξόρυξης γνώσης.

Ο πίνακας METANASTHS_DM, που στον κύβο του Σχήματος 5-2 χρησιμοποιείται ως πίνακας διάστασης αλλά και ως δευτερεύον πίνακας γεγονότων, αποτελεί μια όψη (view) που διατηρεί όλα τα πεδία του πίνακα METANASTHS_OLAP που χρησιμοποιήθηκε στην OLAP ανάλυση, ενώ παράλληλα συγχωνεύει τα πεδία [HPEIROS), [DESC_XVRA_KATAG], [DESC_XVRA_KATAG_ENG] του πίνακα DIM_XWRA_KAT και το πεδίο [OIKOG_KATAST] του πίνακα DIM_OIKOGEN_KATAST. Στο πεδίο [OIKOG_KATAST] χρησιμοποιήθηκε η γενική

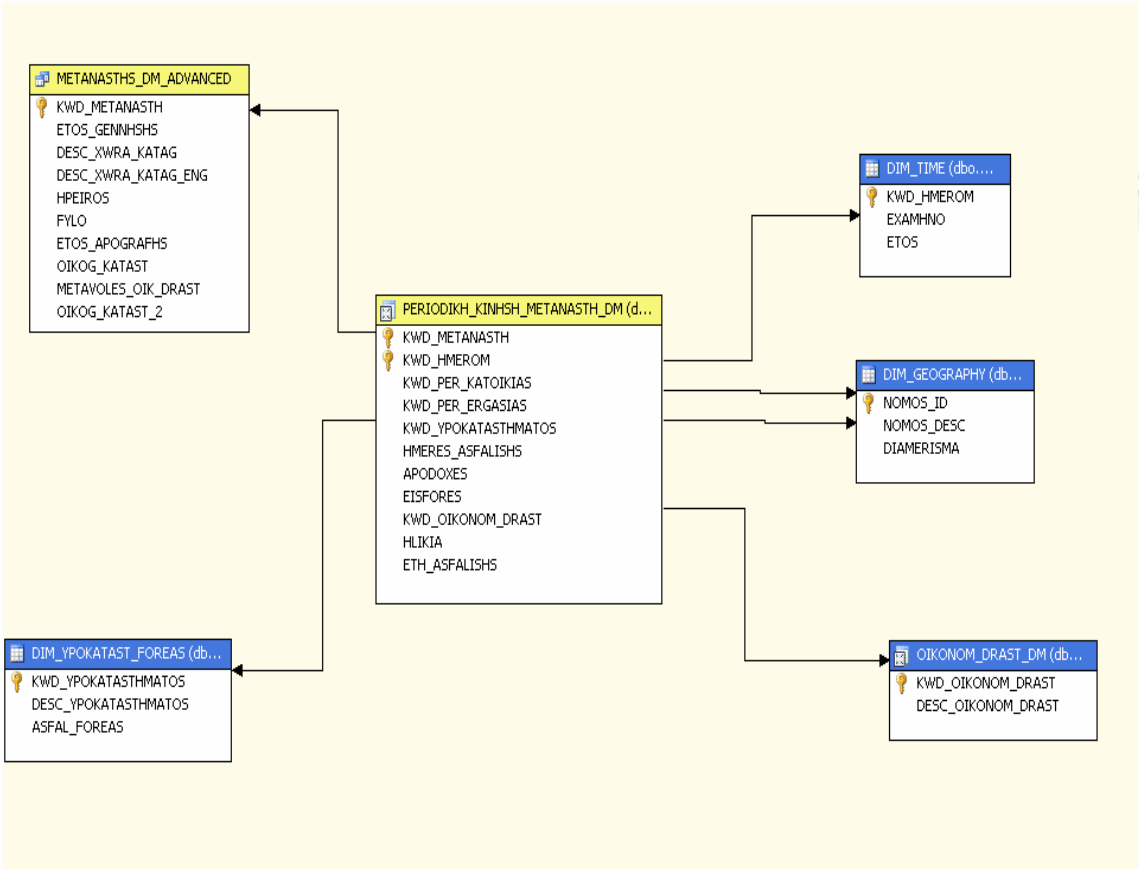
σταθερά “ΑΓΝΩΣΤΟ” ως περιγραφή της οικογενειακής κατάστασης των ασφαλισμένων του ΙΚΑ για τους οποίους δεν υπάρχουν σχετικά διαθέσιμα στοιχεία, η οποία αντικατέστησε τις NULL τιμές. Το SQL Script δημιουργίας του view METANASTHS_DM δίνεται στο Παράρτημα Α.4.

Ο πίνακας METANASTHS_DM_ADVANCED αποτελεί μια διαφορετική μορφή view που ονομάζεται New Named Query (το SQL Script δημιουργίας του δίνεται στο Παράρτημα Α.4) και υλοποιείται στο data source του Analysis Services του Microsoft SQL Server 2005. Ο πίνακας αυτός διατηρεί όλα τα πεδία του view METANASTHS_DM και διαφοροποιείται μόνο ως προς το πεδίο [OIKOG_KATAST], το οποίο δεν συμπληρώθηκε στις κενές του τιμές με χρήση της γενικής σταθεράς “ΑΓΝΩΣΤΟ”. Αντίθετα, έγινε προσπάθεια εφαρμογής μιας άλλης τεχνικής με κάποιες παραλλαγές, ώστε να συμπληρωθεί το πεδίο με πραγματικές τιμές. Σύμφωνα με τη βιβλιογραφία, η διαχείριση ελλιπών τιμών των δεδομένων μπορεί να γίνει και με τη χρήση της μέσης τιμής του χαρακτηριστικού για όλα τα δείγματα που ανήκουν στην ίδια κλάση. Ωστόσο, στην προκειμένη περίπτωση ήταν αδύνατη η χρήση της μέσης τιμής καθώς οι κενές τιμές του πεδίου [OIKOG_KATAST] αναφέρονται σε ένα τελείως διαφορετικό σε σύνθεση σύνολο πληθυσμού, τους ασφαλισμένους αλλοδαπούς του ΙΚΑ_ΕΤΑΜ. Επομένως, αυτό που κάναμε ήταν να αντιστοιχίσουμε τα ποσοστά των ΕΓΓΑΜΩΝ, ΑΓΑΜΩΝ, ΔΙΑΖΕΥΓΜΕΝΩΝ ΚΑΙ ΧΗΡΩΝ (το SQL Script δημιουργίας του view PLITHOS_ANA_OIKOG_KATAST δίνεται στο Παράρτημα Α.4) των ασφαλισμένων του ΟΑΕΕ στον πληθυσμό των ασφαλισμένων του ΙΚΑ. Με τον τρόπο αυτό συμπληρώθηκαν οι κενές τιμές, ενώ παρουσιάστηκε ένα μικρό ποσοστό εγγραφών (περίπου 2%) το οποίο διατήρησε ως τιμή τη γενική σταθερά “ΑΓΝΩΣΤΟ”. Αν και το όλο εγχείρημα απέφερε κάποιο αποτέλεσμα, δεν μπορεί να καταγραφεί ως αποδεκτή λύση στο πρόβλημα διαχείρισης ελλιπών τιμών των δεδομένων και γι’ αυτό δεν χρησιμοποιήθηκε στην περίπτωση των ελλιπών στοιχείων οικονομικής δραστηριότητας για τους ασφαλισμένους του ΟΑΕΕ.

Στη συνέχεια, στα Σχήματα 5-2 και 5-3 παρουσιάζονται οι κύβοι που δημιουργήθηκαν πάνω στο data source του Σχήματος 5-1, στους οποίους χρησιμοποιήθηκαν ως πίνακες γεγονότων και διαστάσεων τα views που περιγράφτηκαν προηγουμένως.



Σχήμα 5-2. Ο κύβος ΙΚΑ_ΟΑΕΕ_DW_SIMPLE.cube



Σχήμα 5-3. Ο κύβος ΙΚΑ_ΟΑΕΕ_DW_ADVANCED.cube

Στους πίνακες που ακολουθούν συνοψίζονται οι διαστάσεις, οι ιεραρχίες και τα μετρήσιμα μεγέθη των κύβων ΙΚΑ_ΟΑΕΕ_DW_SIMPLE.cube και ΙΚΑ_ΟΑΕΕ_DW_ADVANCED.cube, τα οποία εκτός μερικών εξαιρέσεων έχουν σχεδιαστεί με τον ίδιο τρόπο όπως και στην OLAP ανάλυση.

Πίνακας 5-1. Σύνοψη των διαστάσεων και ιεραρχιών των κύβων IKA_OAEE_DW_SIMPLE.cube και IKA_OAEE_DW_ADVANCED.cube

Διάσταση	Ιεραρχία	Περιγραφή
DIM_TIME	**ETOS-EXAMHNO**	Διάσταση Χρόνου
DIM_OIKONOM_DRAST	Δεν έχει οριστεί Ιεραρχία	Διάσταση Οικονομικής Δραστηριότητας των ασφαλισμένων
DIM_YPOKATAST_FOREAS	**ASFALISTIKO TAMEIO** → ASFAL FOREAS	Διάσταση Ασφαλιστικού Φορέα (ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ)
	YPOKATASTHMA → DESC YPOKATASHMATOS	Διάσταση Περιγραφής Υποκαταστημάτων Ασφαλιστικού Φορέα (ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ)
METANASTHS_DM & METANASTHS_DM_ ADVANCED	**PROELEYSH METANASTH** → HPEIROS → ORGANISMOS → DESC_XVRA_KATAG → DESC_XWRA_KATAG_ENG	Διάσταση με δημογραφικά στοιχεία του μετανάστη ασφαλισμένου
	MARITAL STATUS → DESC_OIKOG_KATAST	
	GENDER → **FYLO (Α,Γ)** **METAVOLH ERGASIAS** → METAVOLES OIK DRAST	
DIM_PER_KATOIKIAS	**TOPOTHESIA** → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής κατοικίας του ασφαλισμένου
DIM_PER_ERGASIAS	**TOPOTHESIA** → DIAMERISMA → NOMOS DESC	Διάσταση γεωγραφικής περιοχής εργασίας του ασφαλισμένου
HLIKIAKES OMADES	**AGE_GROUPS** → HLIKIA (διακρίνεται στα διαστήματα) (17-24), (25-29), (30-32), (33-36), (37-40), (41-47) και (48-70)	Διάσταση ηλικιακών ομάδων. Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (DiscretizationMethod - EqualAreas)
PALAIOTHTA	**XRONOS_ASFALISHS** → ETH_ASFALISHS (διακρίνονται στα διαστήματα) (0-3), (4-7), (8-12), (13-16), (17-21), (22-26) και (27-46)	Διάσταση παλαιότητας, δηλαδή χρόνου ασφάλισης κάθε ασφαλισμένου στον αντίστοιχο ασφαλιστικό φορέα. Τα αντίστοιχα διαστήματα προήλθαν με επιλογή από τα Properties (DiscretizationMethod - Clusters)

Πίνακας 5-2. Ομάδες μέτρων (measure groups) των κύβων IKA_OAEE_DW_SIMPLE.cube και IKA_OAEE_DW_ADVANCED.cube

Measure Group	Μετρήσιμα Μεγέθη - Πεδία που δημιουργήθηκαν με τον Cube Wizard	Περιγραφή
PERIODIKH_KINHSH_ METANASTH_DM	PERIODIKH KINHSH METANASTH DM Count	Πλήθος Μεταναστών του συγκεκριμένου πίνακα
	APODOXES	Ποσά αποδοχών ασφαλισμένων (κενό πεδίο)
	EISFORES	Ποσά εισφορών ασφαλισμένων του ΟΑΕΕ
	HLIKIA	Η ηλικία του ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται
	ETH ASFALISHS	Τα έτη ασφάλισης κάθε ασφαλισμένου για κάθε χρονική περίοδο που εξετάζεται
PLITHOS_METANASTWN	UNIQUE METANASTHS	Εύρεση του διακριτού πλήθους των μεταναστών (με count distinct). Βοηθητικό μέτρο.

5.2 Παραδείγματα Δημιουργίας Μοντέλων Εξόρυξης Γνώσης στον Κύβο

Στην προηγούμενη ενότητα ορίστηκαν οι κύβοι δεδομένων με τις διαστάσεις, τις ιεραρχίες και τα μέτρα τους πάνω στους οποίους θα εφαρμοστούν οι τεχνικές εξόρυξης γνώσης. Στη συνέχεια, δίνονται κάποια ενδεικτικά παραδείγματα μοντέλων εξόρυξης γνώσης, τα οποία δημιουργήθηκαν με εφαρμογή των αλγόριθμων Δέντρα Απόφασης (Decision Tress), Συσταδοποίηση (Clustering) και Κανόνες Συσχετίσεων (Association Rules) του εργαλείου Analysis Services του Microsoft SQL Server 2005. Η παρουσίαση και ο σχολιασμός των μοντέλων θα γίνει με τη βοήθεια στιγμιότυπων (screenshot) των αντίστοιχων καρτελών των αλγόριθμων σε κάθε επίπεδο της ανάλυσης. Πρέπει να σημειωθεί ότι η διαδικασία ολοκλήρωσης ενός σεναρίου εξόρυξης γνώσης στο Analysis Services περιλαμβάνει πέντε στάδια: τη δημιουργία της δομής του μοντέλου εξόρυξης γνώσης (mining model structure), τη δημιουργία του μοντέλου εξόρυξης γνώσης (mining model), τη διερεύνηση των μοντέλων εξόρυξης γνώσης, τον έλεγχο της ακρίβειας των μοντέλων εξόρυξης γνώσης (mining accuracy chart) και τέλος, τη δημιουργία προβλέψεων από τα μοντέλα εξόρυξης γνώσης (mining model prediction). Ωστόσο, τα δύο τελευταία στάδια, του ελέγχου της ακρίβειας και της δημιουργίας προβλέψεων, δεν υποστηρίζονται από το σύστημα για μοντέλα που δημιουργούνται πάνω σε κύβους και επομένως στα παραδείγματα που ακολουθούν θα παρουσιάζονται τα αποτελέσματα του τριών πρώτων σταδίων.

5.2.1 Πρόβλεψη της Μεταβολής της Οικονομικής Δραστηριότητας των Ασφαλισμένων Μεταναστών (1η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το predict_metavoles_oik_drast.dmm, το οποίο έχει ως στόχο την πρόβλεψη της μεταβολής της οικονομικής δραστηριότητας των ασφαλισμένων μεταναστών (ανδρών και γυναικών) του ΙΚΑ-ΕΤΑΜ καθώς οι ασφαλισμένοι του

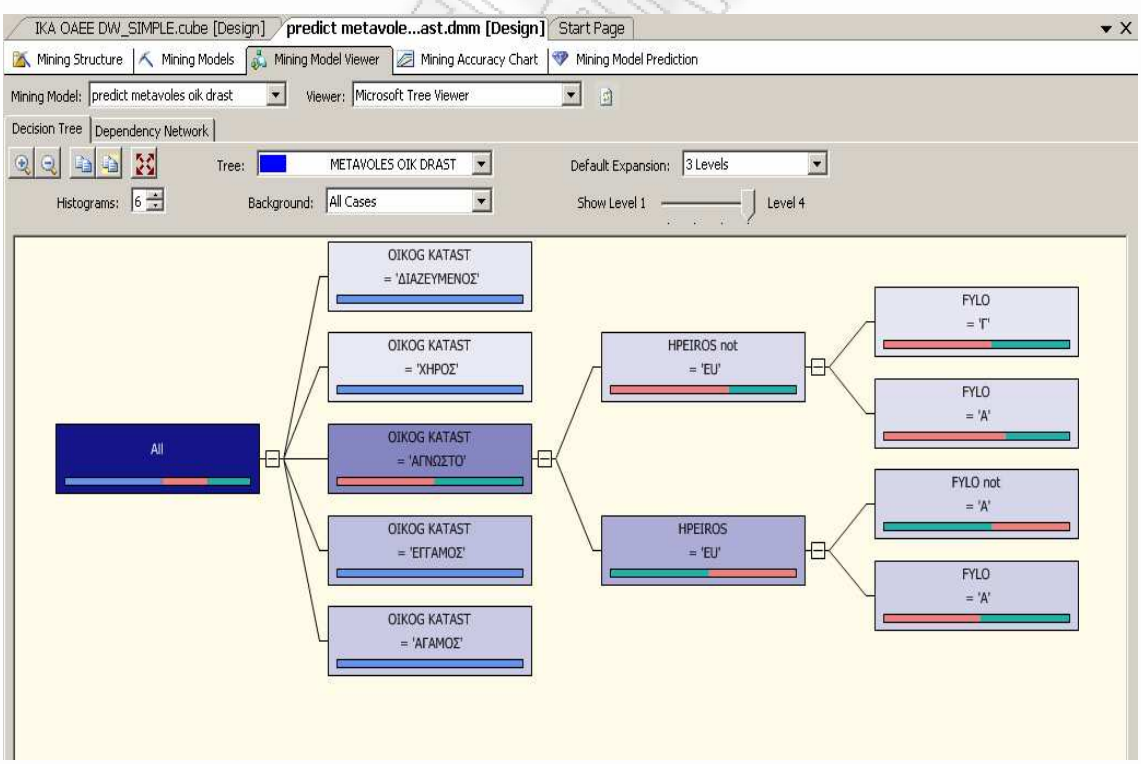
ΟΑΕΕ έχουν εξ' ορισμού σταθερή οικονομική δραστηριότητα, με χρήση του αλγόριθμου Κατηγοριοποίησης Decision Trees του Microsoft SQL Server 2005. Με το συγκεκριμένο μοντέλο δημιουργείται δέντρο απόφασης βάσει του οποίου εντοπίζεται το προφίλ των μεταναστών που έχουν μεγαλύτερη πιθανότητα να αλλάξουν οικονομική δραστηριότητα. Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου ΙΚΑ_ΟΑΕΕ_DW_SIMPLE.cube.

Πίνακας 5-3. Γνωρίσματα του μοντέλου εξόρυξης γνώσης predict_metavoles_oik_drast.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM	METANASTHS_DM	X		
	FYLO		X	
	HPEIROS		X	
	METAVOLES_OIK_DRAST			X
	OIKOG_KATAST		X	
PERIODIKH_KINHSH _METANASTH_DM	HLIKIA		X	
	ETH_ASFALISHS		X	

Επίσης, στον κύβο πραγματοποιήθηκε τεμαχισμός (slice) και ορίστηκαν τα φίλτρα χρονική περίοδος {1^{ο} εξάμηνο του 2008} και επιλογή μεταναστών από τα γεωγραφικά διαμερίσματα {Στερεά Ελλάδα, Πελοπόννησος, Μακεδονία}, στα αντίστοιχα πεδία των πινάκων διαστάσεων DIM_TIME και DIM_PER_ERGASIAS.

Στη συνέχεια, μετά την ολοκλήρωση της δημιουργίας του μοντέλου που ορίστηκε με βάση τα παραπάνω γνωρίσματα και φίλτρα, ακολουθεί η διερεύνηση του μοντέλου εξόρυξης γνώσης.

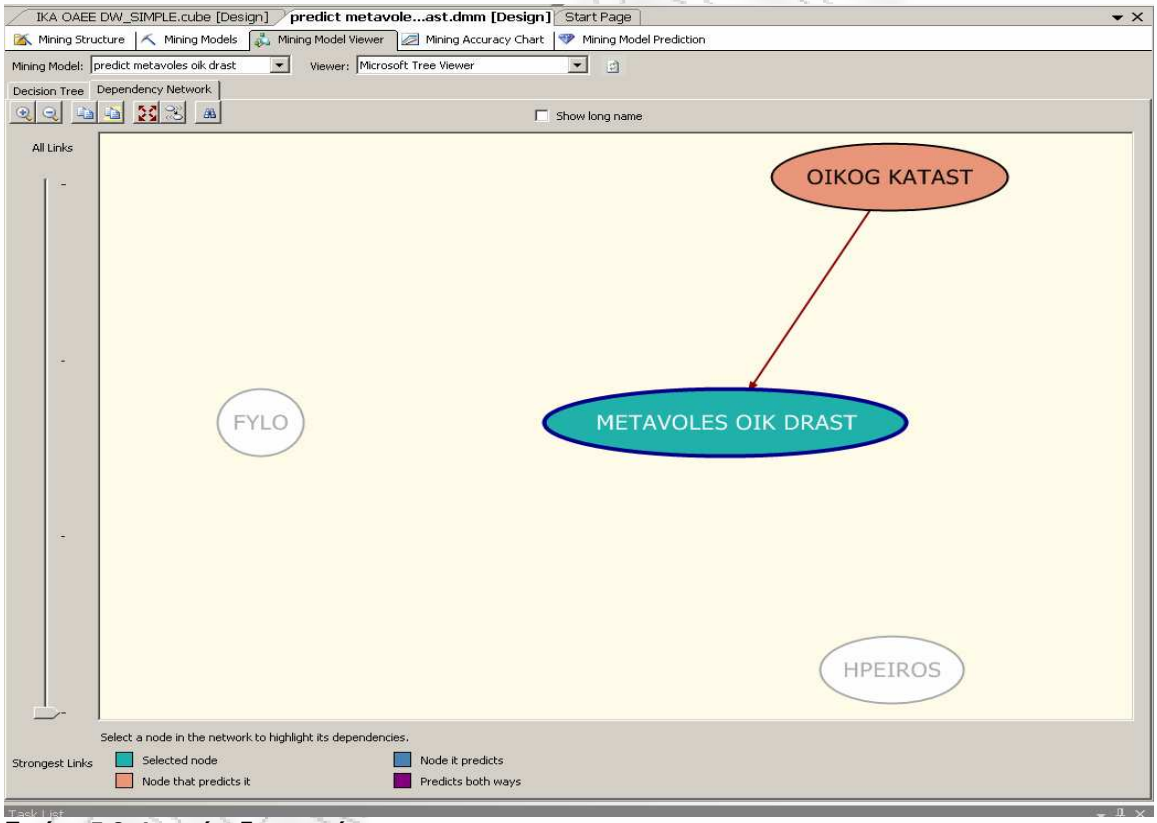


Εικόνα 5-1. Το εξαγόμενο δέντρο απόφασης

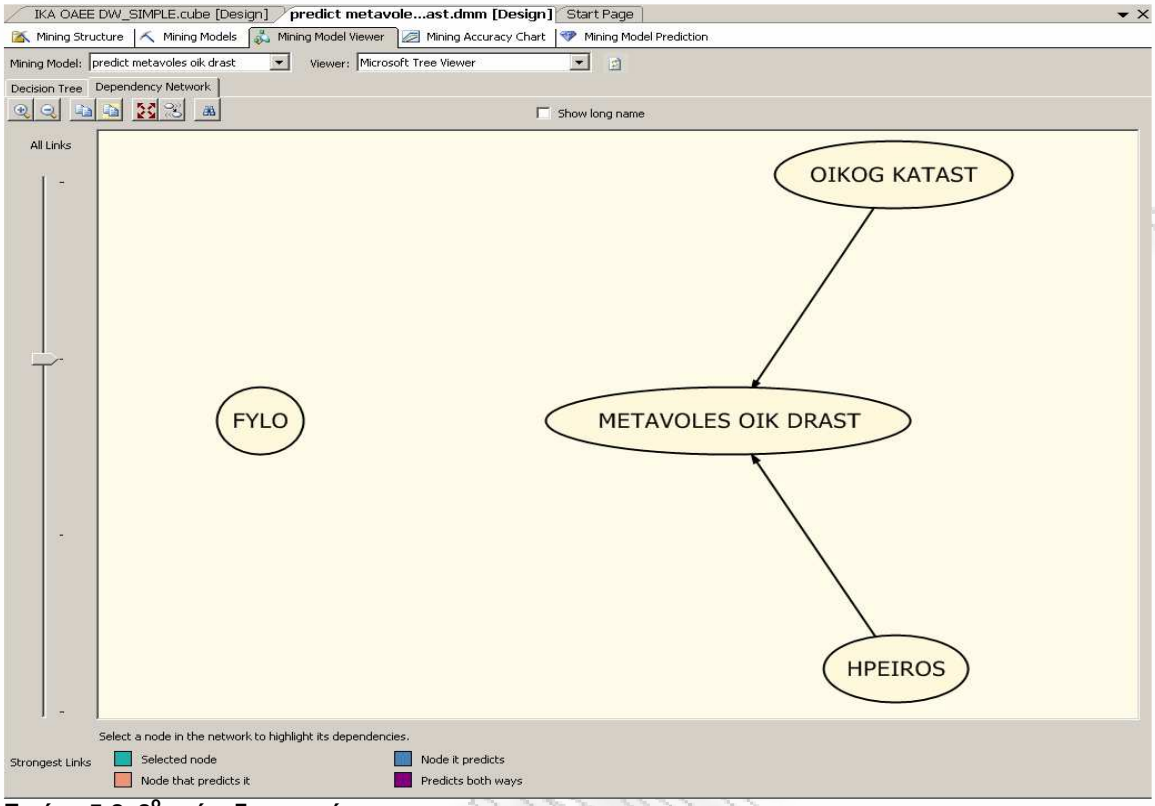
Όπως παρατηρούμε στο παραπάνω δέντρο απόφασης ο αλγόριθμος δίνει αποτελέσματα για τον πληθυσμό των ασφαλισμένων του ΙΚΑ-ΕΤΑΜ που βρίσκεται στο δεύτερο επίπεδο κόμβων με άγνωστη την οικογενειακή κατάσταση. Οι υπόλοιποι κόμβοι του δευτέρου επιπέδου

δεν αναλύονται περαιτέρω καθώς αναφέρονται στους ασφαλισμένους του ΟΑΕΕ των οποίων η οικονομική δραστηριότητα δεν μεταβάλλεται. Βάσει των αποτελεσμάτων του αλγορίθμου σε σύνολο 14444 περιπτώσεων οι Άνδρες ασφαλισμένοι του ΙΚΑ, που δεν προέρχονται από χώρες τις Ευρώπης διατηρούν σταθερή την οικονομική τους δραστηριότητα με ποσοστό 65,04% έναντι του ποσοστού 34,96% που αλλάζουν. Τα αντίστοιχα ποσοστά για αυτούς που προέρχονται από χώρες της Ευρώπης είναι 51,67% με σταθερή οικονομική δραστηριότητα και 48,33% με οικονομική δραστηριότητα που αλλάζει σε σύνολο 26095 περιπτώσεων. Όσον αφορά στις Γυναίκες ασφαλισμένες του ΙΚΑ που δεν προέρχονται από χώρες της Ευρώπης το πρότυπο διατηρείται, καθώς σε σύνολο 3677 περιπτώσεων διατηρούν σταθερή την οικονομική τους δραστηριότητα με ποσοστό 57,24% έναντι του ποσοστού 42,75% αυτών που αλλάζουν. Τα αντίστοιχα ποσοστά για αυτές που προέρχονται από χώρες της Ευρώπης είναι 42,22% με σταθερή οικονομική δραστηριότητα και 57,78% με οικονομική δραστηριότητα που αλλάζει σε σύνολο 21268 περιπτώσεων. Επομένως, οι αλλοδαποί ασφαλισμένοι που προέρχονται από άλλες χώρες πλην των χωρών της Ευρώπης τείνουν να διατηρούν σταθερή την οικονομική τους δραστηριότητα σε σχέση με τους προερχόμενους από χώρες της Ευρώπης.

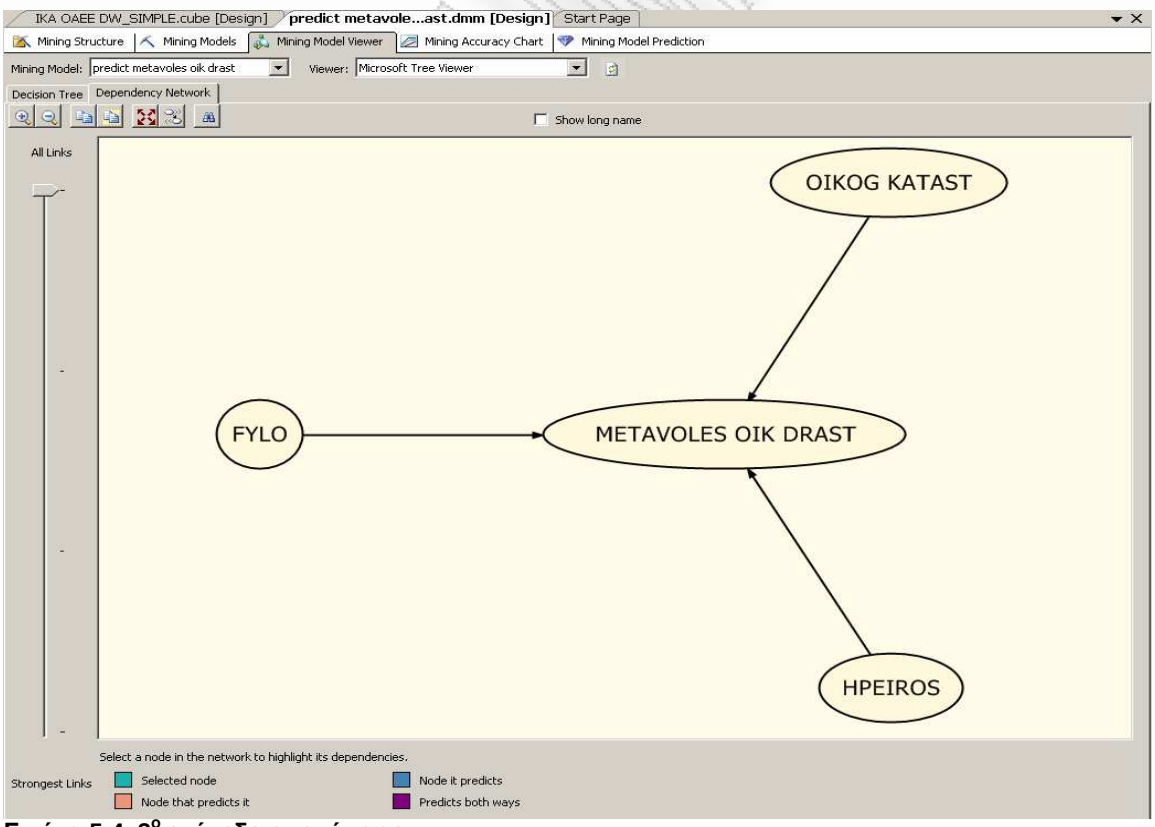
Συνεχίζοντας τη διερεύνηση του μοντέλου, επιλέγουμε την καρτέλα “Dependency Network” στην οποία απεικονίζονται οι συσχετίσεις των γνωρισμάτων εισόδου με το γνώρισμα προς πρόβλεψη. Τα εξαγόμενα αποτελέσματα δίνονται στις παρακάτω εικόνες.



Εικόνα 5-2. 1ο επίπεδο συσχέτισης



Εικόνα 5-3. 2^{ο} επίπεδο συσχέτισης



Εικόνα 5-4. 3^{ο} επίπεδο συσχέτισης

Όπως παρατηρούμε στην καρτέλα “Dependency Network”, δεν απεικονίζονται όλα τα αρχικά γνωρίσματα εισόδου, παρά μόνο τα πιο ισχυρά για την εξαγωγή αποτελεσμάτων τα

οποία απεικονίστηκαν και στο δέντρο απόφασης. Η Εικόνα 5-2 απεικονίζει την πιο ισχυρή συσχέτιση μεταξύ γνωρίσματος εισόδου και γνωρίσματος προς πρόβλεψη, και ακολουθούν τα επόμενα επίπεδα συσχέτισης γνωρισμάτων των Εικόνων 5-3 και 5-4.

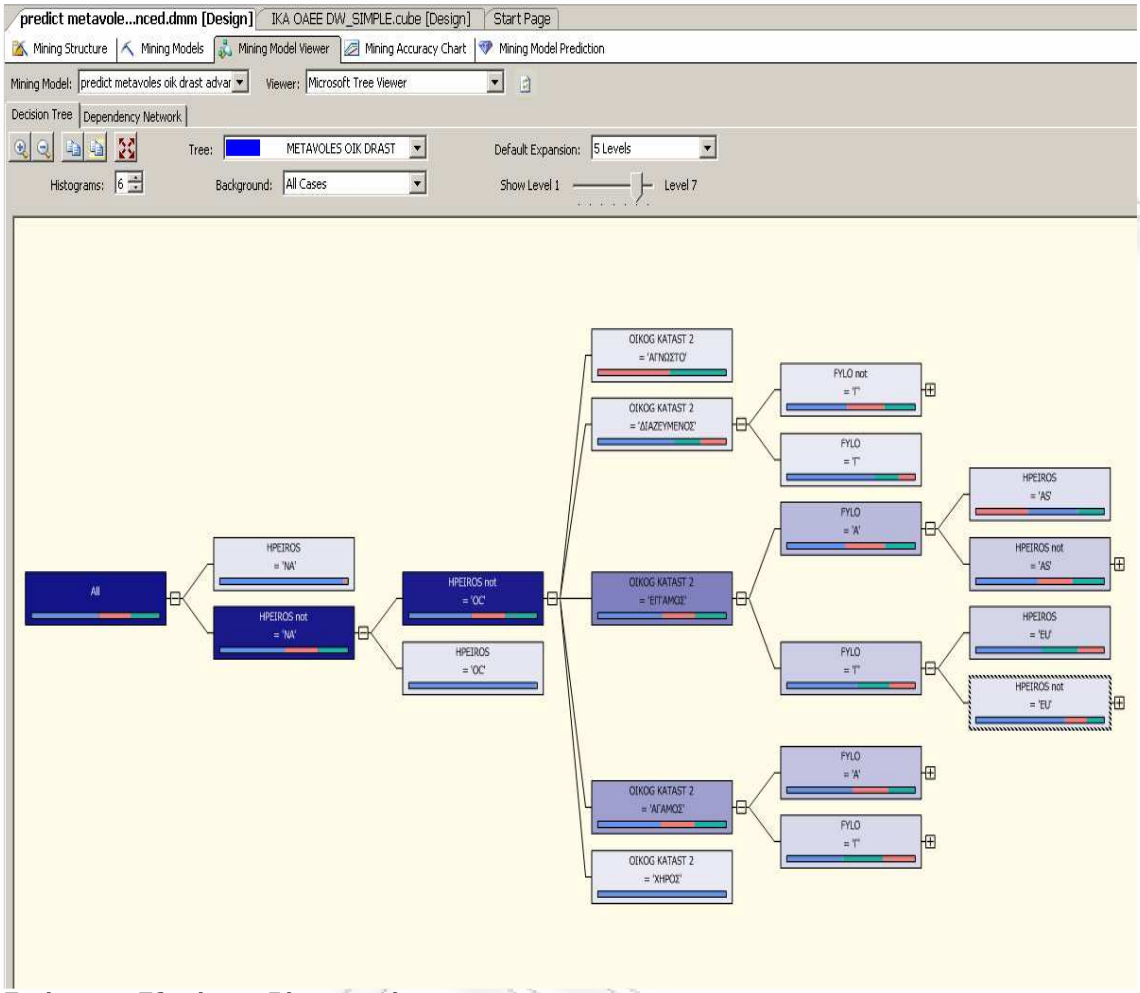
5.2.2 Πρόβλεψη της Μεταβολής της Οικονομικής Δραστηριότητας των Ασφαλισμένων Μεταναστών (2η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το predict_metavoles_oik_drast_ advanced.dmm, το οποίο έχει ως στόχο την πρόβλεψη της μεταβολής της οικονομικής δραστηριότητας των ασφαλισμένων μεταναστών (ανδρών και γυναικών), με χρήση του αλγορίθμου Κατηγοριοποίησης Decision Trees του Microsoft SQL Server 2005. Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου ΙΚΑ_ΟΑΕΕ_DW_ADVANCED.cube. Στο συγκεκριμένο κύβο είχαν συμπληρωθεί στοιχεία οικογενειακής κατάστασης και για τους ασφαλισμένους του ΙΚΑ-ΕΤΑΜ τα οποία θα χρησιμοποιηθούν ώστε να γίνει αντιπαραβολή με τα αποτελέσματα του παραδείγματος της παραγράφου 5.2.1.

Πίνακας 5-4. Γνωρίσματα του μοντέλου εξόρυξης γνώσης predict_metavoles_oik_drast_ advanced.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM_ ADVANCED	METANASTHS_DM_AD VANCED	X		
	FYLO		X	
	HPEIROS		X	
	METAVOLES_OIK_DRA ST			X
	OIKOG_KATAST_2		X	
HLIKIAKES_OMAD ES	PERIODIKH_KINHSH_ METANASTH _DM	X		
	HLIKIA		X	
PALAIOTHTA	PERIODIKH_KINHSH_ METANASTH _DM	X		
	ETH_ASFALISHS		X	

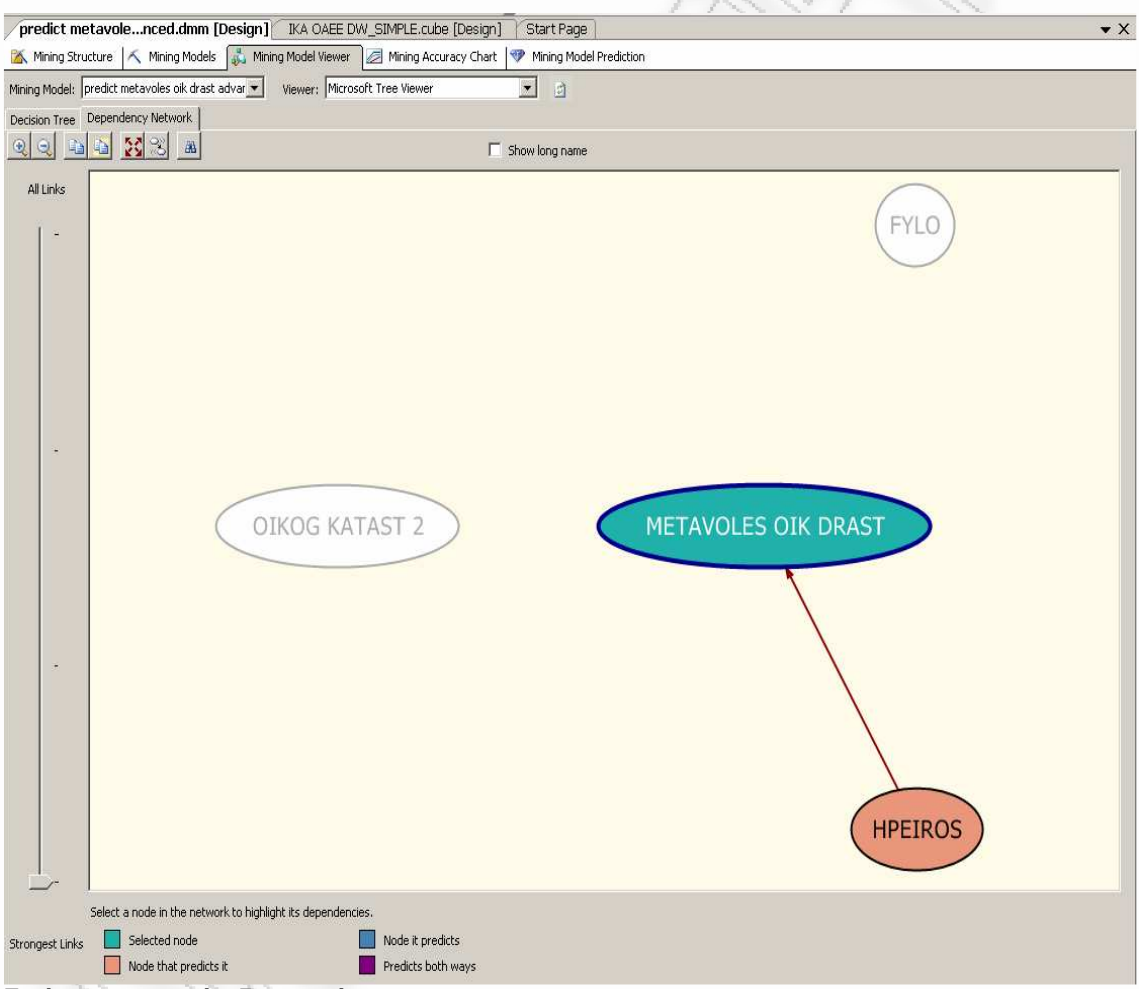
Επειδή στο μοντέλο predict_metavoles_oik_drast_ advanced.dmm εφαρμόστηκαν τα ίδια φίλτρα με αυτά στη παραγράφου 5.2.1 συνεχίζουμε με τη διερεύνησή του.



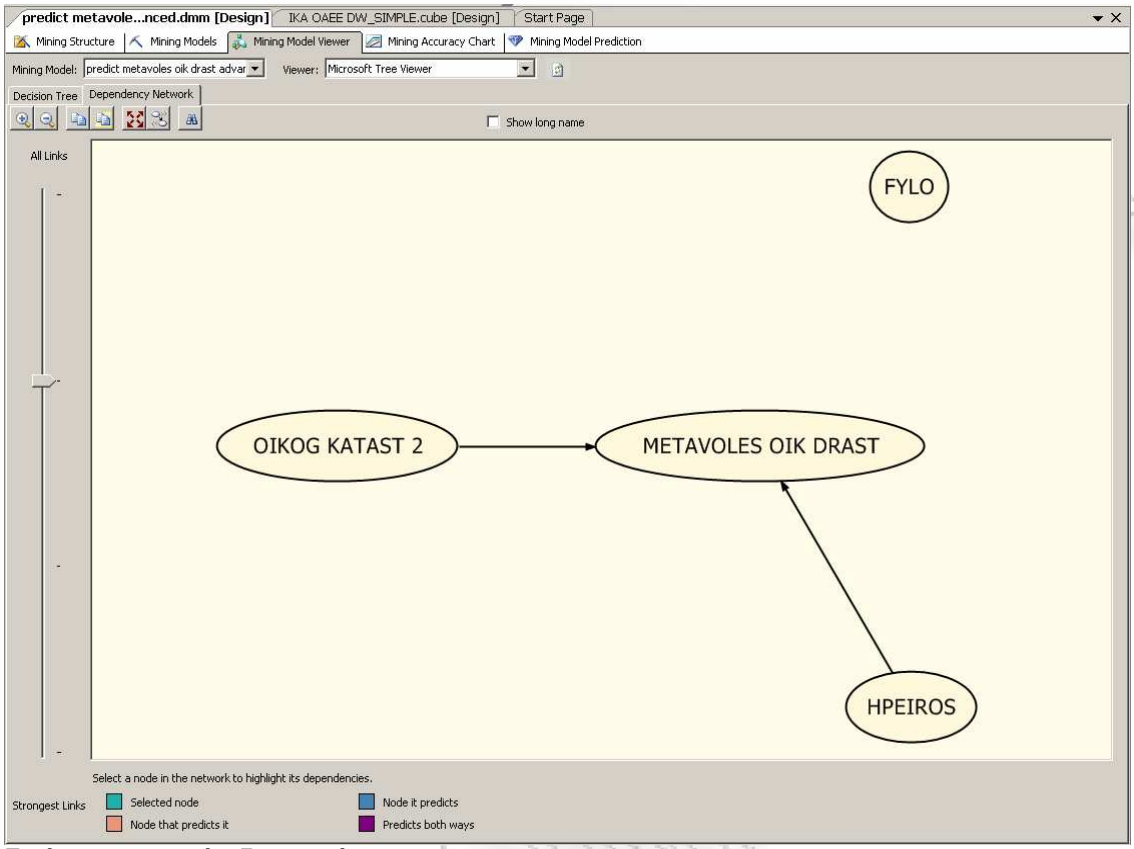
Εικόνα 5-5. Εξαγόμενο δέντρο απόφασης

Όπως παρατηρούμε στην παραπάνω εικόνα, το δέντρο απόφασης που δίνει ο αλγόριθμος είναι αρκετά πιο πολύπλοκο (επτά συνολικά επίπεδα ανάπτυξης, έξι φαίνονται στην εικόνα) από αυτό της προηγούμενης παραγράφου. Μια πρώτη παρατήρηση στο δέντρο απόφασης είναι ότι αποκλείονται στο δεύτερο κιόλας επίπεδο οι ασφαλισμένοι που προέρχονται από τις ηπείρους Βόρεια Αμερική (NA) και Ωκεανία (OC), κυρίως λόγω του μικρού αριθμού ασφαλισμένων που προέρχονται από αυτές. Στη συνέχεια, για λόγους συντομίας, θα σχολιαστούν ενδεικτικά μονοπάτια του δέντρου απόφασης καθώς η πλήρης ανάλυση δεν προσδίδει επιπλέον συμπεράσματα. Επιλέγουμε το μονοπάτι HPEIROS not = 'NA' και HPEIROS not = 'OC' και OIKOG KATAST 2 = 'ΕΓΓΑΜΟΣ' και FYLO = 'Α' και HPEIROS = 'AS', στο οποίο σε σύνολο 11558 περιπτώσεων οι Άνδρες ασφαλισμένοι, που προέρχονται από χώρες τις Ασίας διατηρούν σταθερή την οικονομική τους δραστηριότητα με ποσοστό 39,83% έναντι του ποσοστού 20,82% αυτών που αλλάζουν. Τα αντίστοιχα ποσοστά για αυτούς που δεν προέρχονται από χώρες της Ασίας (μονοπάτι HPEIROS not = 'NA' και HPEIROS not = 'OC' και OIKOG KATAST 2 = 'ΕΓΓΑΜΟΣ' και FYLO = 'Α' και HPEIROS not = 'AS') είναι 27,07% με σταθερή οικονομική δραστηριότητα και 24,90% με οικονομική δραστηριότητα που αλλάζει σε σύνολο 28395 περιπτώσεων. Όσον αφορά στις Γυναίκες ασφαλισμένες που δεν προέρχονται από χώρες της Ευρώπης, ενδεικτικά επιλέγουμε το μονοπάτι HPEIROS not = 'NA' και HPEIROS not = 'OC' και OIKOG KATAST 2 = 'ΕΓΓΑΜΟΣ' και FYLO = 'Γ' και HPEIROS not= 'EU', βάσει του οποίου σε σύνολο 5798 περιπτώσεων αυτές που διατηρούν σταθερή την οικονομική τους δραστηριότητα έχουν ποσοστό 17,54% έναντι του ποσοστού 13,56% αυτών που αλλάζουν. Τα αντίστοιχα ποσοστά για αυτές που προέρχονται από χώρες της Ευρώπης (μονοπάτι HPEIROS not = 'NA' και HPEIROS not = 'OC' και OIKOG KATAST 2 = 'ΕΓΓΑΜΟΣ' και FYLO = 'Γ' και HPEIROS = 'EU') είναι 20,75% με σταθερή οικονομική δραστηριότητα και 28,78% με οικονομική δραστηριότητα που αλλάζει σε σύνολο 22140 περιπτώσεων.

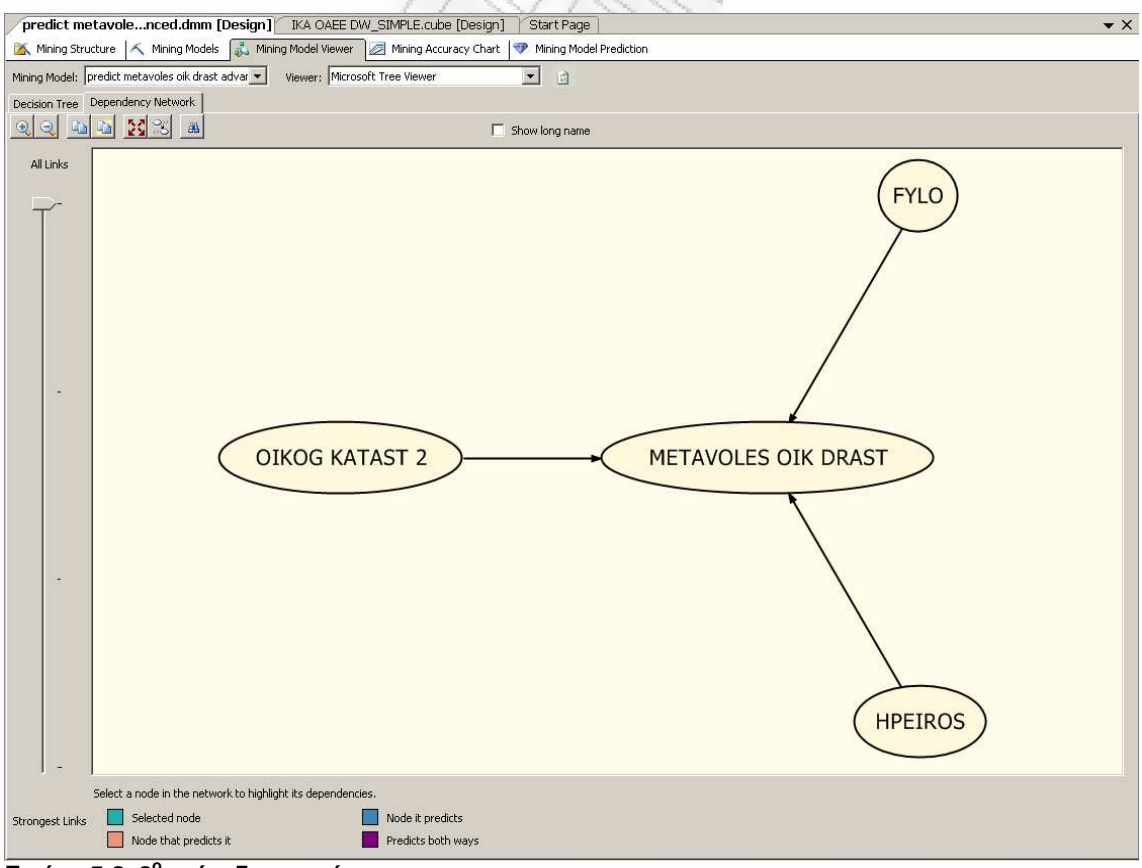
Επομένως, παρόλο που οι ασφαλισμένοι του ΙΚΑ απέκτησαν στοιχεία οικογενειακής κατάστασης ώστε να μη διαφοροποιούνται από τους ασφαλισμένους του ΟΑΕΕ το τελικό αποτέλεσμα, δηλαδή η γνώση στην οποία καταλήγουμε είναι η ίδια με αυτή της προηγούμενης παραγράφου. Δηλαδή, οι αλλοδαποί ασφαλισμένοι που προέρχονται από άλλες χώρες πλην των χωρών της Ευρώπης τείνουν να διατηρούν σταθερή την οικονομική τους δραστηριότητα σε σχέση με τους προερχόμενους από χώρες της Ευρώπης. Το συμπέρασμα αυτό είναι λογικό καθώς μεταβολές οικονομικής δραστηριότητας παρουσιάζουν μόνο οι ασφαλισμένοι του ΙΚΑ οπότε ο αλγόριθμος έδωσε παρόμοια αποτελέσματα. Το μόνο που μεταβλήθηκε με την παρέμβαση στα στοιχεία της οικογενειακής κατάστασης των ασφαλισμένων του ΙΚΑ, όπως φαίνεται στα παρακάτω screenshot των επιπέδων της καρτέλας “Dependency Network”, είναι το γνώρισμα της πιο ισχυρής συσχέτισης. Στην προηγούμενη παράγραφο η μεταβολή της οικονομικής δραστηριότητας σχετιζόταν ισχυρά με την οικογενειακή κατάστάση (γνώρισμα OIKOG_KATAST) ενώ στην εξεταζόμενη περίπτωση σχετίζεται με την ήπειρο προέλευσης του ασφαλισμένου μετανάστη (γνώρισμα HPEIROS).



Εικόνα 5-6. 1ο επίπεδο συσχέτισης



Εικόνα 5-7. 2ο επίπεδο συσχέτισης



Εικόνα 5-8. 3^{ο} επίπεδο συσχέτισης

5.2.3 Συσταδοποίηση Ασφαλισμένων Μεταναστών με Βάση τα Δημογραφικά τους Στοιχεία (1η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το clustering_metanasths_ demographics.dmm, το οποίο έχει ως στόχο την ομαδοποίηση των ασφαλισμένων μεταναστών σε συστάδες με παρόμοια χαρακτηριστικά με βάση τα δημογραφικά τους στοιχεία και την ανακάλυψη σχέσεων μεταξύ των χαρακτηριστικών κάθε συστάδας. Η υλοποίηση του μοντέλου πραγματοποιήθηκε με χρήση του αλγορίθμου Συσταδοποίησης Clustering του Microsoft SQL Server 2005.

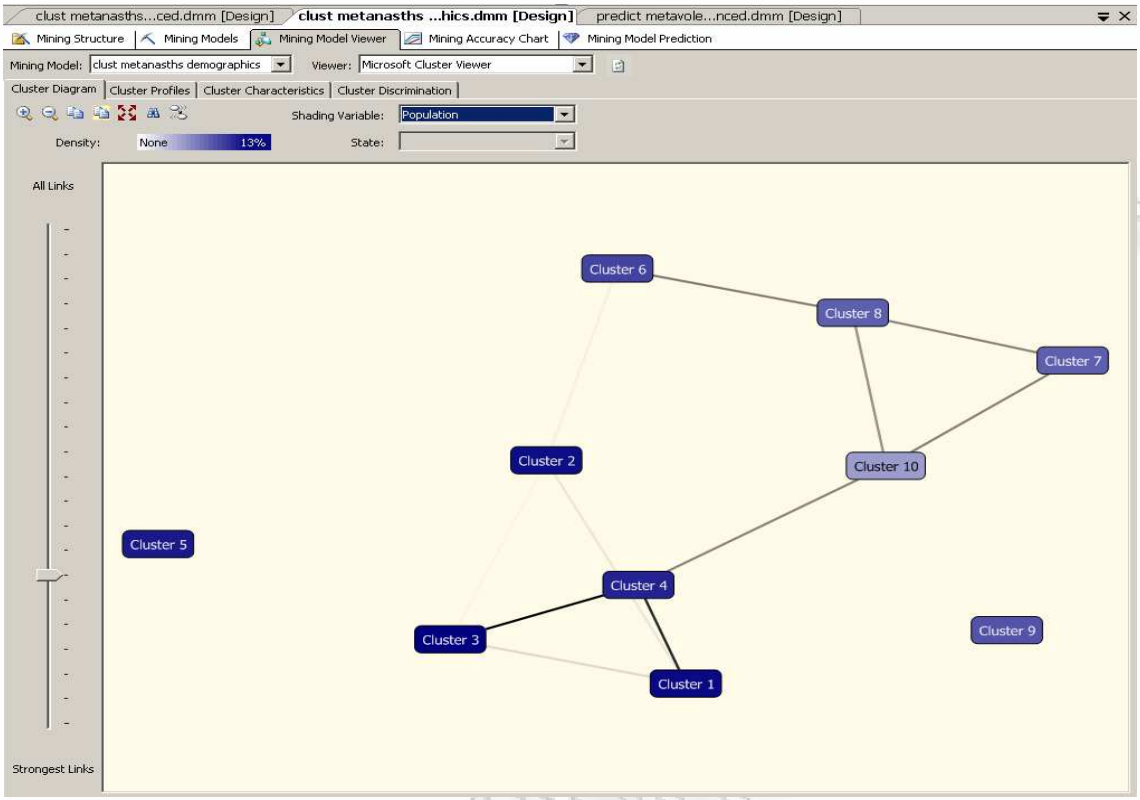
Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου IKA_OAEE_DW_SIMPLE.cube.

Πίνακας 5-5. Γνωρίσματα του μοντέλου εξόρυξης γνώσης clustering_metanasths_ demographics.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM	METANASTHS_DM	X		
	FYLO		X	
	HPEIROS		X	
PERIODIKH_KINHSH _METANASTH_DM	HLIKIA		X	X
	ETH_ASFALISHS		X	

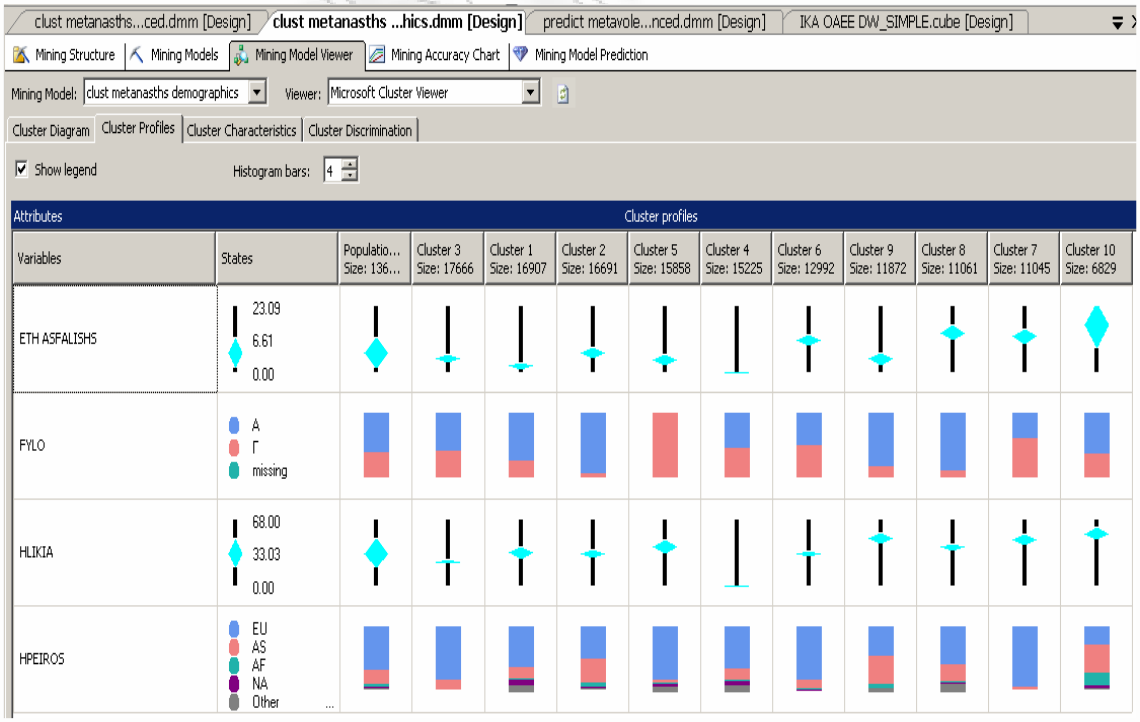
Επίσης, στον κύβο πραγματοποιήθηκε τεμαχισμός (slice) και ορίστηκε ως φίλτρο η χρονική περίοδος εξετάζεται να είναι το {1^{ο} εξάμηνο του 2007}, αντίστοιχο πεδίο του πίνακα DIM_TIME.

Στη συνέχεια, μετά την ολοκλήρωση της δημιουργίας του μοντέλου που ορίστηκε με βάση τα παραπάνω γνωρίσματα και φίλτρα, ακολουθεί η διερεύνηση του μοντέλου εξόρυξης γνώσης μελετώντας στιγμιότυπα των καρτελών Cluster Diagram (διάγραμμα συσχετίσεων συστάδων), Cluster Profiles (προφίλ συστάδων), Cluster Characteristics (χαρακτηριστικά συστάδων) και Cluster Discrimination (διαφοροποίηση συστάδων).



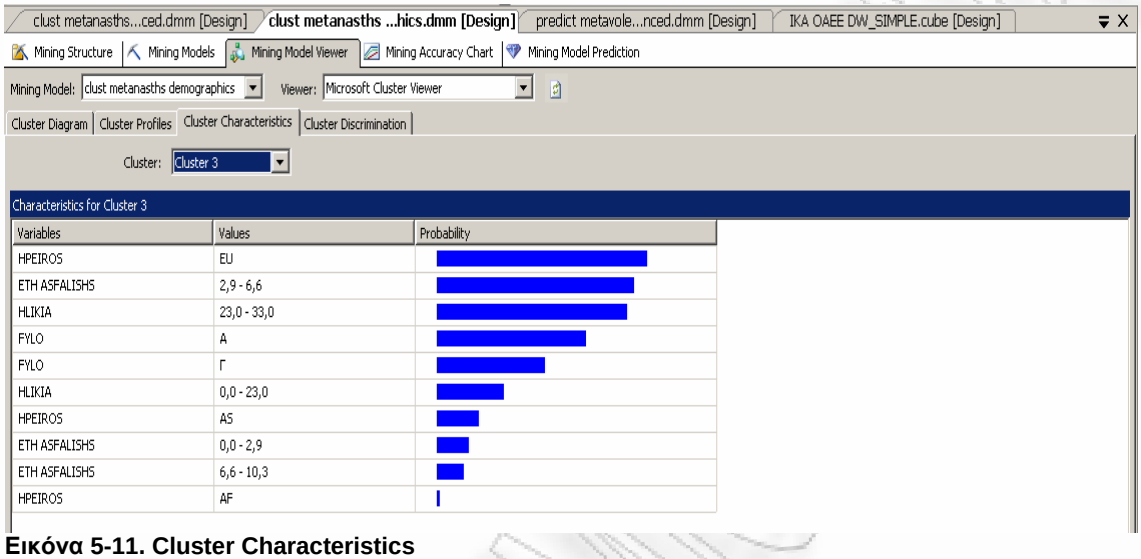
Εικόνα 5-9. Cluster Diagram

Όπως παρατηρούμε στην παραπάνω εικόνα, το διάγραμμα συσχετίσεων που δημιουργήθηκε αποτελείται από δέκα συστάδες (βάσει της default τιμής της παραμέτρου Cluster_Count του αλγορίθμου) και οι συστάδες 1, 4 και 3, 4 παρουσιάζουν ισχυρές σχέσεις μεταξύ τους. Επίσης, οι συστάδες 1,3 λόγω του έντονου χρωματισμού τους, αποτελούν τις συστάδες με το μεγαλύτερο αριθμό εγγραφών στο σύνολο του πληθυσμού.



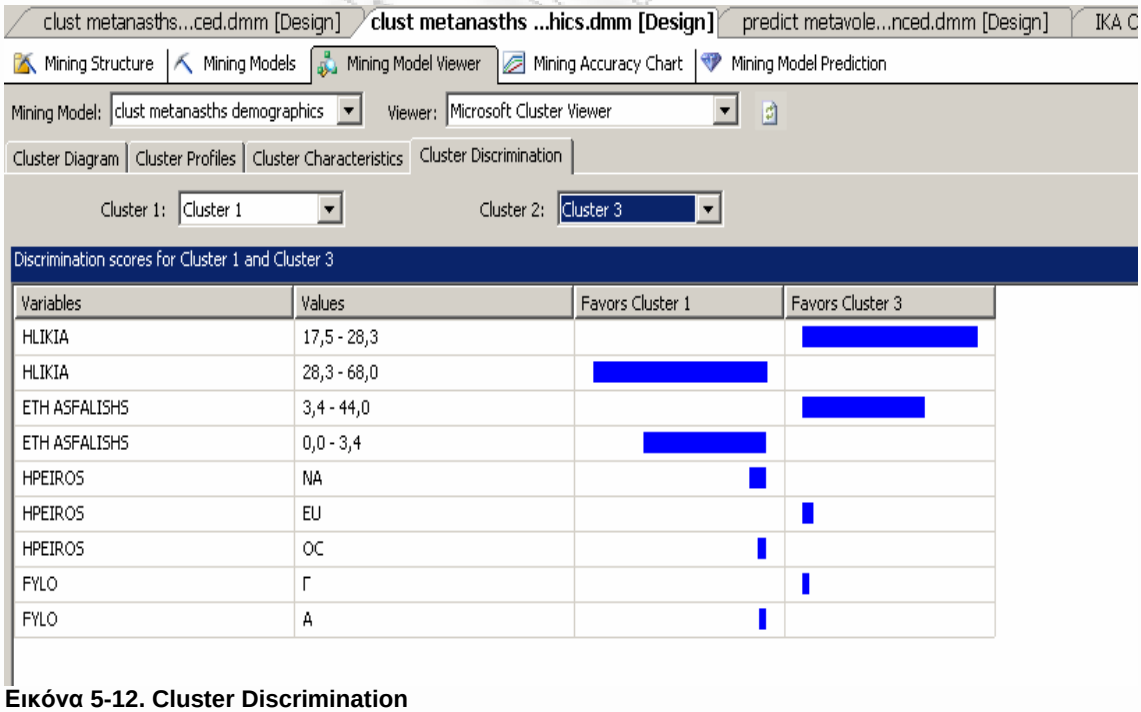
Εικόνα 5-10. Cluster Profiles

Η Εικόνα 5-10 απεικονίζει μια γενική θεώρηση του συγκεκριμένου μοντέλου συσταδοποίησης, δηλαδή το σύνολο των συστάδων με αναλογίες τιμών για τα γνωρίσματά τους. Τα διακριτά γνωρίσματα εισόδου παρουσιάζονται με χρωματιστές ράβδους ενώ τα συνεχή γνωρίσματα εισόδου με διαγράμματα διαμαντιού (diamond chart), που αντιπροσωπεύουν το μέσο όρο και την τυπική απόκλιση σε κάθε συστάδα.



Εικόνα 5-11. Cluster Characteristics

Στην καρτέλα Cluster Characteristics εξετάζονται τα χαρακτηριστικά κάθε συστάδας με μεγαλύτερη λεπτομέρεια. Τα σχετικά γνωρίσματα ταξινομούνται σε φθίνουσα σειρά. Επομένως, βάσει της Εικόνας 5-11, η συστάδα 3 με τον μεγαλύτερο αριθμό εγγραφών περιλαμβάνει κυρίως άνδρες ασφαλισμένους που προέρχονται από χώρες της Ευρώπης, ηλικίας μεταξύ 23-33 ετών, των οποίων τα έτη ασφάλισης στους ασφαλιστικούς φορείς κυμαίνονται μεταξύ 2,9 και 6,6 ετών.



Εικόνα 5-12. Cluster Discrimination

Στην καρτέλα Cluster Discrimination εξετάζονται τα χαρακτηριστικά που διαφοροποιούν μια συστάδα από μια άλλη. Στην Εικόνα 5-12 γίνεται σύγκριση μεταξύ των συστάδων 1 και 3. Όπως παρατηρούμε η συστάδα 1 περιλαμβάνει άνδρες ασφαλισμένους, μικρούς σε ηλικία (μεταξύ 17,5-28,3 ετών), με λίγα έτη ασφάλισης (από 0 έως 3,4 έτη) που προέρχονται από χώρες της Βόρειας Αμερικής. Αντίθετα, η συστάδα 3 περιλαμβάνει γυναίκες ασφαλισμένες, ηλικίας μεταξύ 28,3 και 68 ετών, με μεγάλο εύρος ετών ασφάλισης (από 3,4 έως 44 έτη) που προέρχονται από χώρες της Ευρώπης. Επομένως, δεν υπάρχει μεγάλη συσχέτιση μεταξύ των συστάδων 1 και 3 καθώς τα χαρακτηριστικά τους διαφέρουν σημαντικά, γεγονός που απεικονίζεται και στο αρχικό διάγραμμα συσχετίσεων.

5.2.4 Συσταδοποίηση Ασφαλισμένων Μεταναστών με Βάση τα Δημογραφικά τους Στοιχεία (2η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το clustering_metanasths_ profile_advanced.dmm, το οποίο έχει ως στόχο την ομαδοποίηση των ασφαλισμένων μεταναστών σε συστάδες με παρόμοια χαρακτηριστικά με βάση τα δημογραφικά τους στοιχεία και την ανακάλυψη σχέσεων μεταξύ των χαρακτηριστικών κάθε συστάδας. Η υλοποίηση του μοντέλου πραγματοποιήθηκε με χρήση του αλγόριθμου Συσταδοποίησης Clustering του Microsoft SQL Server 2005.

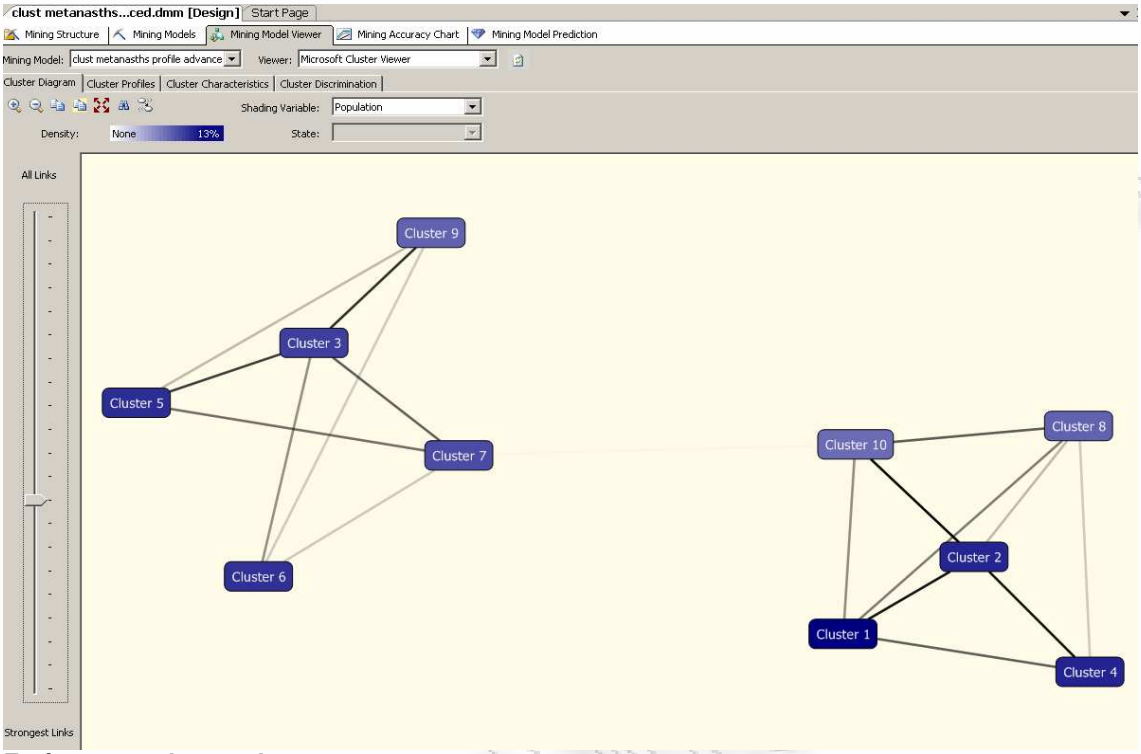
Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου ΙΚΑ_ΟΑΕΕ_DW_ADVANCED.cube. Ουσιαστικά, με την προσάρτηση επιπλέον γνωρισμάτων στον αλγόριθμο (χώρα καταγωγής, οικογενειακή κατάσταση μεταναστών του ΙΚΑ και του ΟΑΕΕ, μεταβολή της οικονομικής τους δραστηριότητας) επιθυμούμε να μελετήσουμε τις σχέσεις που αναπτύσσονται μέσα στις διάφορες συστάδες και συνθέτουν τα αντίστοιχα προφίλ των μεταναστών.

Πίνακας 5-6. Γνωρίσματα του μοντέλου εξόρυξης γνώσης clustering_metanasths_ profile_advanced.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM_ ADVANCED	METANASTHS_DM_ ADVANCED	X		
	FYLO		X	
	DESC_XVRA_KATAG		X	
	METAVOLES_OIK_ DRAST		X	
	OIKOG_KATAST_2		X	
PERIODIKH_KINHSH _METANASTH_DM	HLIKIA		X	X
	ETH_ASFALISHS		X	

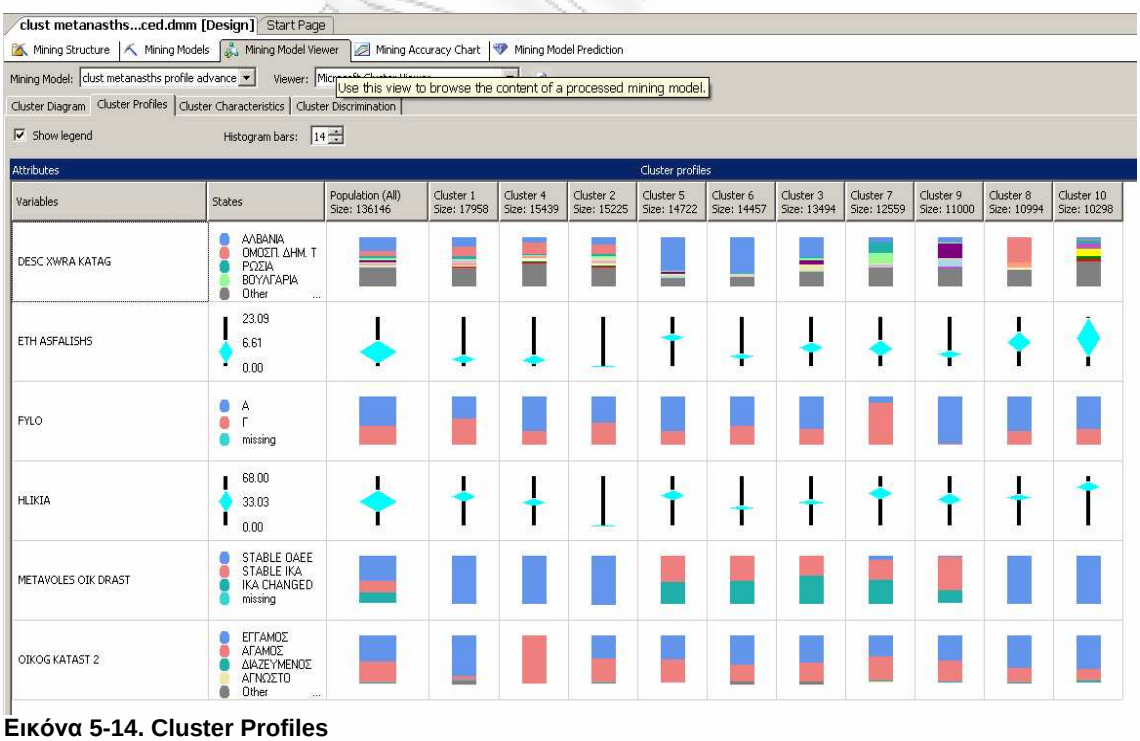
Επίσης, στον κύβο πραγματοποιήθηκε τεμαχισμός (slice) και ορίστηκε ως φίλτρο η χρονική περίοδος που εξετάζεται να είναι το {1^{ο} εξάμηνο του 2007}, αντίστοιχο πεδίο του πίνακα DIM_TIME, που μαζί με τον ορισμό των γνωρισμάτων ολοκληρώνει τη διαδικασία δημιουργίας του μοντέλου.

Στη συνέχεια, ακολουθεί η διερεύνηση του μοντέλου εξόρυξης γνώσης μελετώντας στιγμιότυπα των καρτελών Cluster Diagram (διάγραμμα συσχετίσεων συστάδων), Cluster Profiles (προφίλ συστάδων), Cluster Characteristics (χαρακτηριστικά συστάδων) και Cluster Discrimination (διαφοροποίηση συστάδων).



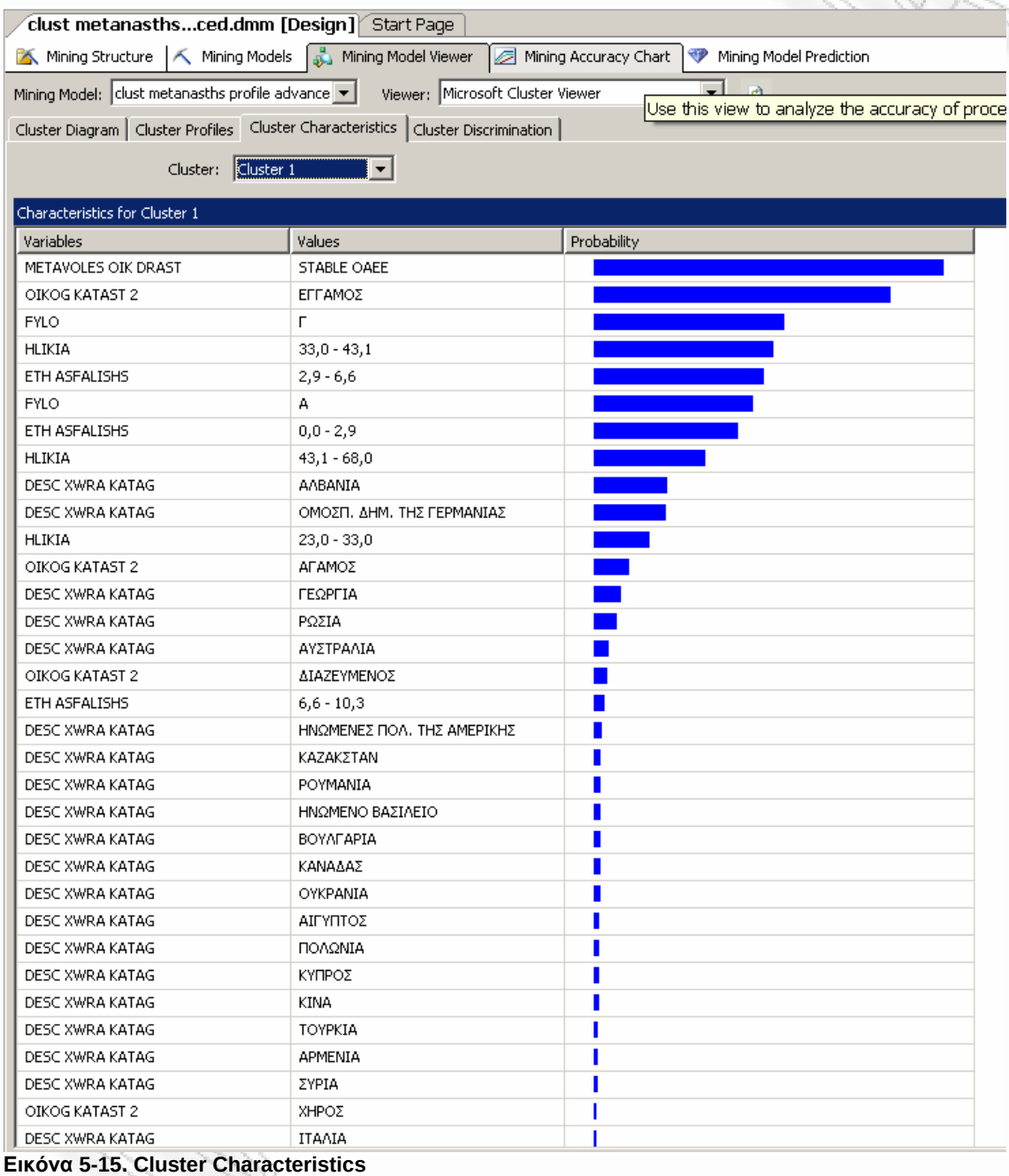
Εικόνα 5-13. Cluster Diagram

Όπως παρατηρούμε στην παραπάνω εικόνα, το διάγραμμα συσχετίσεων που δημιουργήθηκε αποτελείται από δέκα συστάδες (βάσει της default τιμής της παραμέτρου Cluster_Count του αλγορίθμου) και οι συστάδες 1, 2 και 2,4 παρουσιάζουν ισχυρές σχέσεις μεταξύ τους. Επίσης, οι συστάδες 1,4 λόγω του έντονου χρωματισμού τους, αποτελούν τις συστάδες με το μεγαλύτερο αριθμό εγγραφών στο σύνολο του πληθυσμού.



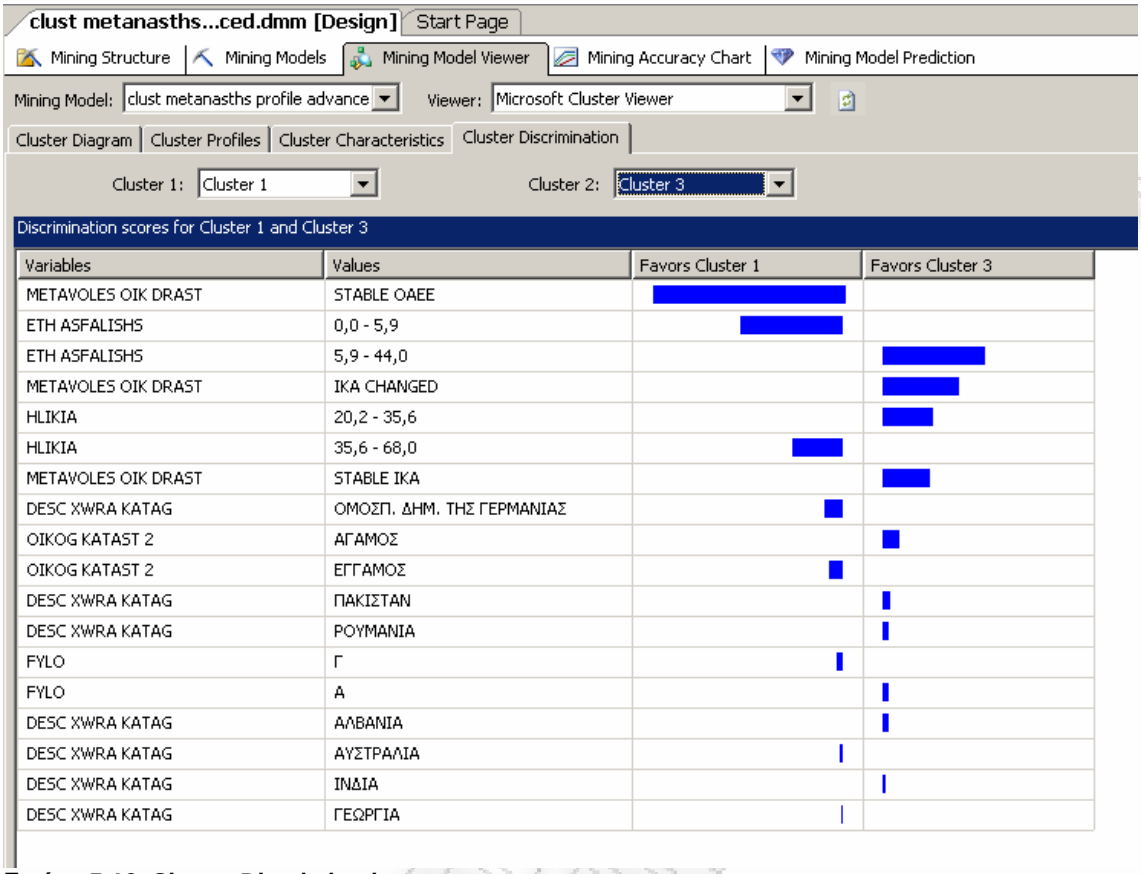
Εικόνα 5-14. Cluster Profiles

Η παραπάνω Εικόνα απεικονίζει μια γενική θεώρηση του συγκεκριμένου μοντέλου συσταδοποίησης, δηλαδή το σύνολο των συστάδων με αναλογίες τιμών για τα γνωρίσματά τους. Τα διακριτά γνωρίσματα εισόδου παρουσιάζονται με χρωματιστές ράβδους ενώ τα συνεχή γνωρίσματα εισόδου με διαγράμματα διαμαντιού (diamond chart), που αντιπροσωπεύουν το μέσο όρο και την τυπική απόκλιση σε κάθε συστάδα.



Εικόνα 5-15. Cluster Characteristics

Στην καρτέλα Cluster Characteristics εξετάζονται τα χαρακτηριστικά κάθε συστάδας με μεγαλύτερη λεπτομέρεια. Τα σχετικά γνωρίσματα ταξινομούνται σε φθίνουσα σειρά. Επομένως, βάσει της Εικόνας 5-15, η συστάδα 1 με τον μεγαλύτερο αριθμό εγγραφών περιλαμβάνει κυρίως έγγαμες γυναίκες ασφαλισμένες του ΟΑΕΕ που προέρχονται από την Αλβανία, ηλικίας μεταξύ 33-43 ετών, οι οποίες ασφαλίζονται στον συγκεκριμένο ασφαλιστικό φορέα για μικρό χρονικό διάστημα που κυμαίνεται μεταξύ 2,9 και 6,6 ετών.



Variables	Values	Favors Cluster 1	Favors Cluster 3
METAVOLES OIK DRAST	STABLE OAEE		
ETH ASFALISHS	0,0 - 5,9		
ETH ASFALISHS	5,9 - 44,0		
METAVOLES OIK DRAST	IKA CHANGED		
HLIKIA	20,2 - 35,6		
HLIKIA	35,6 - 68,0		
METAVOLES OIK DRAST	STABLE IKA		
DESC XWRA KATAG	ΟΜΟΣΠ. ΔΗΜ. ΤΗΣ ΓΕΡΜΑΝΙΑΣ		
OIKOG KATAST 2	ΑΓΑΜΟΣ		
OIKOG KATAST 2	ΕΓΓΑΜΟΣ		
DESC XWRA KATAG	ΠΑΚΙΣΤΑΝ		
DESC XWRA KATAG	ΡΟΥΜΑΝΙΑ		
FYLO	Γ		
FYLO	Α		
DESC XWRA KATAG	ΑΛΒΑΝΙΑ		
DESC XWRA KATAG	ΑΥΣΤΡΑΛΙΑ		
DESC XWRA KATAG	ΙΝΔΙΑ		
DESC XWRA KATAG	ΓΕΩΡΓΙΑ		

Εικόνα 5-16. Cluster Discrimination

Στην καρτέλα Cluster Discrimination εξετάζονται τα χαρακτηριστικά που διαφοροποιούν μια συστάδα από μια άλλη. Στην Εικόνα 5-16 γίνεται σύγκριση μεταξύ των συστάδων 1 και 3. Όπως παρατηρούμε η συστάδα 1 περιλαμβάνει έγγαμες γυναίκες ασφαλισμένες του ΟΑΕΕ, ηλικίας μεταξύ 35,6 και 68 ετών, με λίγα έτη ασφάλισης που κυμαίνονται από 0 έως 5,9 έτη και προέρχονται από την Ομόσπονδη Δημοκρατία της Γερμανίας. Αντίθετα, η συστάδα 3 περιλαμβάνει άγαμους άνδρες ασφαλισμένους του ΙΚΑ, με μεταβαλλόμενη οικονομική δραστηριότητα, ηλικίας μεταξύ 20,2 και 35,6 ετών, με μεγάλο εύρος ετών ασφάλισης (από 5,9 έως 44 έτη) που προέρχονται από το Πακιστάν και τη Ρουμανία. Επομένως, δεν υπάρχει μεγάλη συσχέτιση μεταξύ των συστάδων 1 και 3 καθώς τα χαρακτηριστικά τους διαφέρουν σημαντικά, γεγονός που απεικονίζεται και στο αρχικό διάγραμμα συσχετίσεων.

5.2.5 Εύρεση Κανόνων Συσχετίσεων Μεταξύ Δημογραφικών Δεδομένων Ασφαλισμένων Μεταναστών (1η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το metanasths_ association_rules1.dmm, το οποίο έχει ως στόχο την εύρεση συχνών συνόλων αντικειμένων με τα δημογραφικά δεδομένα των ασφαλισμένων μεταναστών στους ασφαλιστικούς οργανισμούς ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ, τους πιθανούς κανόνες συσχέτισης των αντικειμένων, το ποσοστό υποστήριξης και το επίπεδο εμπιστοσύνης. Η υλοποίηση του μοντέλου πραγματοποιήθηκε με χρήση του αλγόριθμου Κανόνων Συσχετίσεων Association Rules του Microsoft SQL Server 2005.

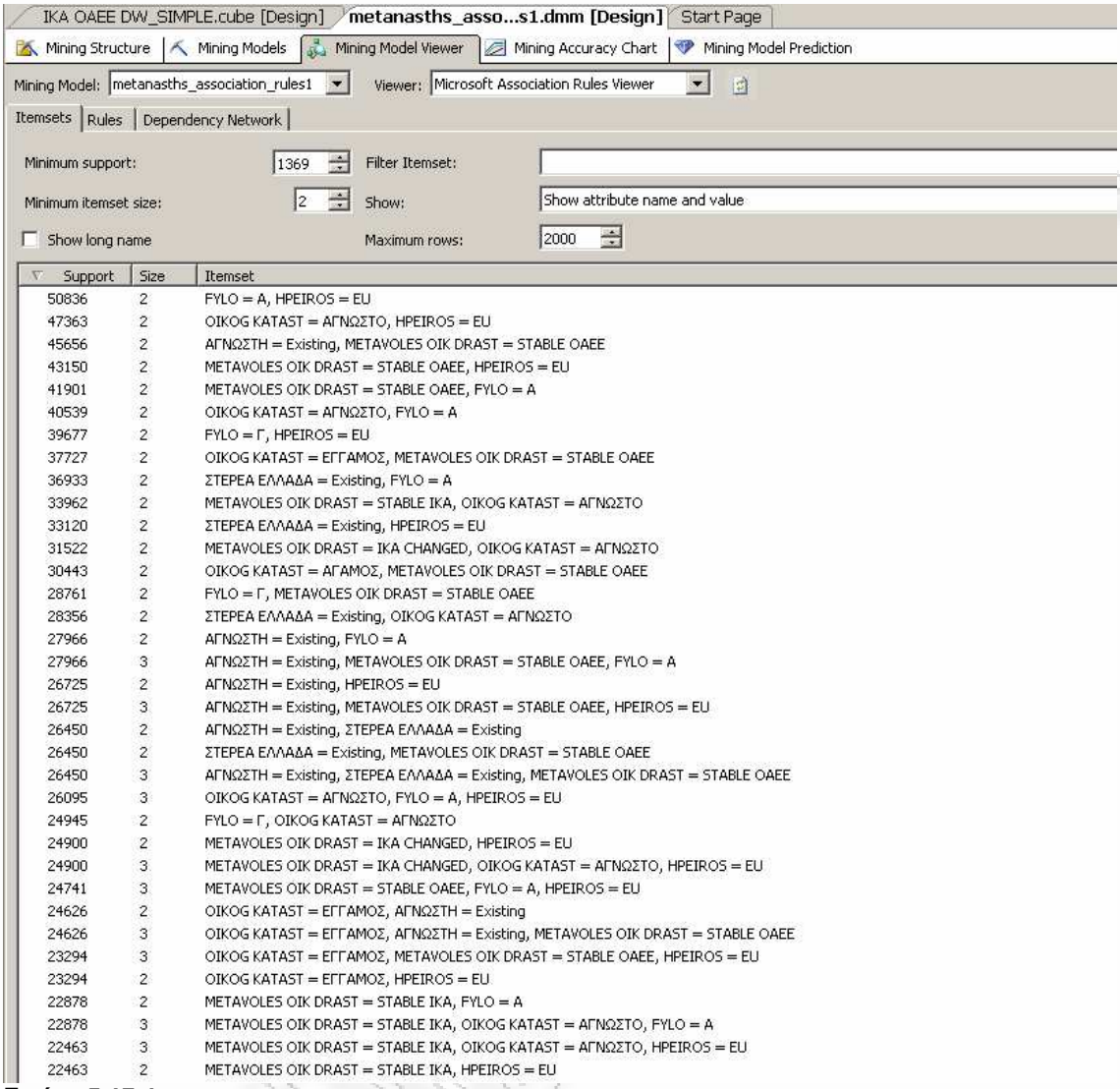
Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου ΙΚΑ_ΟΑΕΕ_DW_SIMPLE.cube.

Πίνακας 5-7. Γνωρίσματα του μοντέλου εξόρυξης γνώσης metanasths_ association_rules1.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM	METANASTHS_DM	**X**		
	FYLO		**X**	
	HPEIROS		**X**	
	METAVOLES_OIK_ DRAST		**X**	
	OIKOG_KATAST		**X**	
OIKONOM_DRAST_ DM	DESC_OIKONOM_ DRAST	**X**	**X**	
PER_KATOIKIAS	DIAMERISMA	**X**		**X**

Επίσης, στον κύβο πραγματοποιήθηκε τεμαχισμός (slice) και ορίστηκαν, από τα αντίστοιχα πεδία των πινάκων διαστάσεων DIM_TIME,DIM_PER_KATOIKIAS και OIKONOM_DRAST_ DM, τα εξής φίλτρα: χρονική περίοδος που εξετάζεται είναι το {1^{ο} εξάμηνο του 2008}, γεωγραφικό διαμέρισμα που διαμένουν οι ασφαλισμένοι μετανάστες κάποιο από τα {ΣΤΕΡΕΑ ΕΛΛΑΔΑ, ΠΕΛΟΠΟΝΝΗΣΟΣ, ΜΑΚΕΔΟΝΙΑ} και κατηγορίες οικονομικής δραστηριότητας των μεταναστών κάποια από τις {ΑΛΙΕΙΑ, ΓΕΩΡΓΙΑ,ΚΤΗΝΟΤΡΟΦΙΑ,ΘΗΡΑ ΚΑΙ ΔΑΣΟΚΟΜΙΑ, ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ, ΟΡΥΧΕΙΑ ΚΑΙ ΛΑΤΟΜΕΙΑ, ΚΑΤΑΣΚΕΥΕΣ, ΜΕΤΑΦΟΡΕΣ,ΑΠΟΘΗΚΕΥΣΗ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΕΣ, ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ, ΑΓΝΩΣΤΗ}. Επίσης, στις παραμέτρους του αλγορίθμου MINIMUM_PROBABILITY (ελάχιστη πιθανότητα) και MINIMUM_SUPPORT (ελάχιστη υποστήριξη) θέσαμε τις τιμές 0,1 και 0,01 αντίστοιχα. Η παράμετρος MINIMUM_PROBABILITY καθορίζει την ελάχιστη πιθανότητα με την οποία ένας κανόνας συσχέτισης είναι αληθής ενώ η παράμετρος MINIMUM_SUPPORT καθορίζει τον μέγιστο αριθμό περιπτώσεων βάσει των οποίων υποστηρίζεται ένα στοιχειοσύνολο. Ο καθορισμός φίλτρων και τιμών παραμέτρων μαζί με τον ορισμό των γνωρισμάτων ολοκληρώνει τη διαδικασία δημιουργίας του μοντέλου.

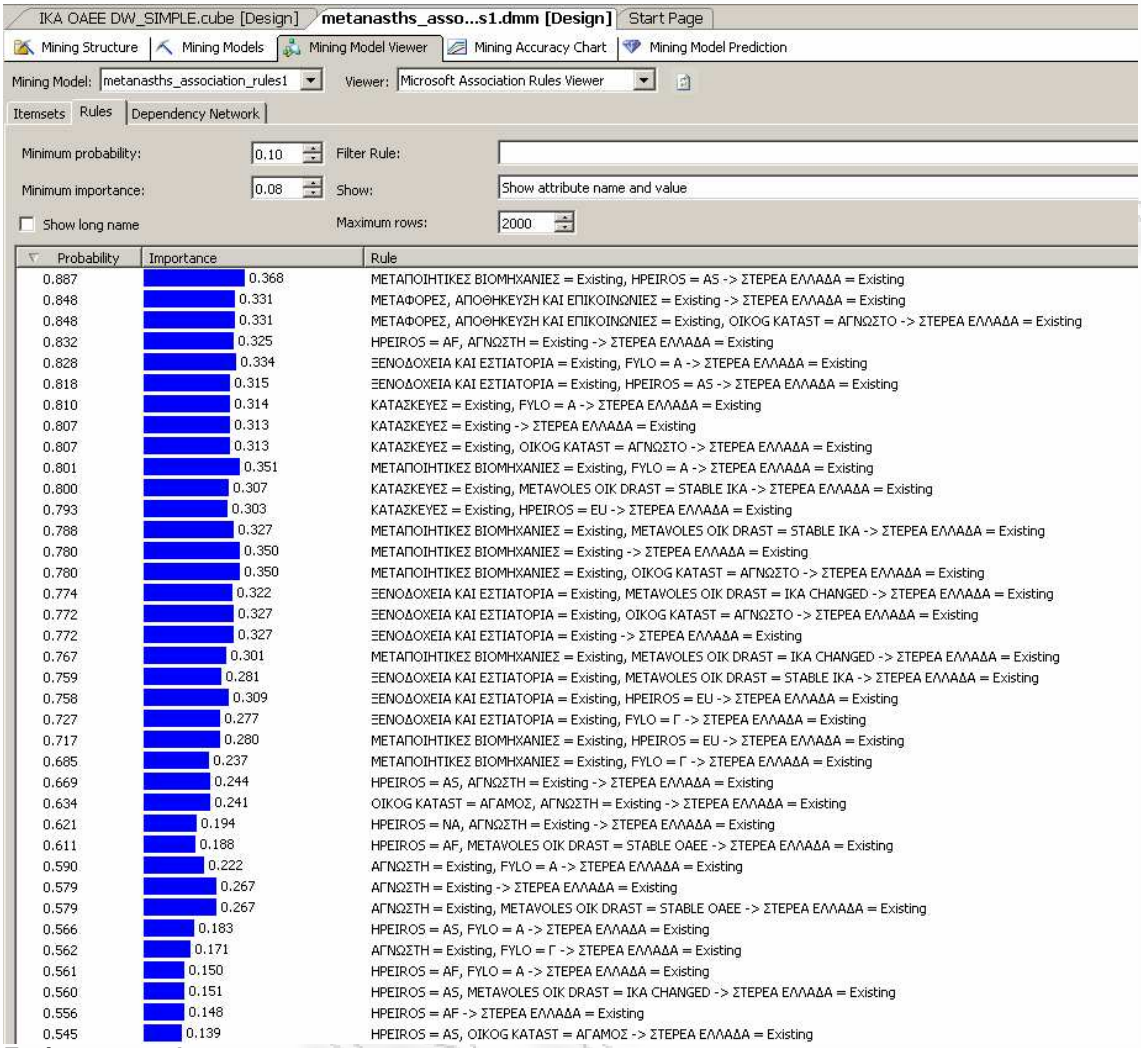
Στη συνέχεια, ακολουθεί η διερεύνηση του μοντέλου εξόρυξης γνώσης μελετώντας στιγμιότυπα των καρτελών Itemsets (στοιχειοσύνολα), Rules (κανόνες) και Dependency Network (διάγραμμα συσχετίσεων).



Support	Size	Itemset
50836	2	FYLO = A, HPEIROS = EU
47363	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, HPEIROS = EU
45656	2	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE
43150	2	METAVOLES OIK DRAST = STABLE OAEE, HPEIROS = EU
41901	2	METAVOLES OIK DRAST = STABLE OAEE, FYLO = A
40539	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, FYLO = A
39677	2	FYLO = Γ, HPEIROS = EU
37727	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΕΓΓΑΜΟΣ, METAVOLES OIK DRAST = STABLE OAEE
36933	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, FYLO = A
33962	2	METAVOLES OIK DRAST = STABLE IKA, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ
33120	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, HPEIROS = EU
31522	2	METAVOLES OIK DRAST = IKA CHANGED, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ
30443	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΑΜΟΣ, METAVOLES OIK DRAST = STABLE OAEE
28761	2	FYLO = Γ, METAVOLES OIK DRAST = STABLE OAEE
28356	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ
27966	2	ΑΓΝΩΣΤΗ = Existing, FYLO = A
27966	3	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE, FYLO = A
26725	2	ΑΓΝΩΣΤΗ = Existing, HPEIROS = EU
26725	3	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE, HPEIROS = EU
26450	2	ΑΓΝΩΣΤΗ = Existing, ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing
26450	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, METAVOLES OIK DRAST = STABLE OAEE
26450	3	ΑΓΝΩΣΤΗ = Existing, ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, METAVOLES OIK DRAST = STABLE OAEE
26095	3	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, FYLO = A, HPEIROS = EU
24945	2	FYLO = Γ, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ
24900	2	METAVOLES OIK DRAST = IKA CHANGED, HPEIROS = EU
24900	3	METAVOLES OIK DRAST = IKA CHANGED, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, HPEIROS = EU
24741	3	METAVOLES OIK DRAST = STABLE OAEE, FYLO = A, HPEIROS = EU
24626	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΕΓΓΑΜΟΣ, ΑΓΝΩΣΤΗ = Existing
24626	3	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΕΓΓΑΜΟΣ, ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE
23294	3	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΕΓΓΑΜΟΣ, METAVOLES OIK DRAST = STABLE OAEE, HPEIROS = EU
23294	2	ΟΙΚΟΓ ΚΑΤΑΣΤ = ΕΓΓΑΜΟΣ, HPEIROS = EU
22878	2	METAVOLES OIK DRAST = STABLE IKA, FYLO = A
22878	3	METAVOLES OIK DRAST = STABLE IKA, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, FYLO = A
22463	3	METAVOLES OIK DRAST = STABLE IKA, ΟΙΚΟΓ ΚΑΤΑΣΤ = ΑΓΝΩΣΤΟ, HPEIROS = EU
22463	2	METAVOLES OIK DRAST = STABLE IKA, HPEIROS = EU

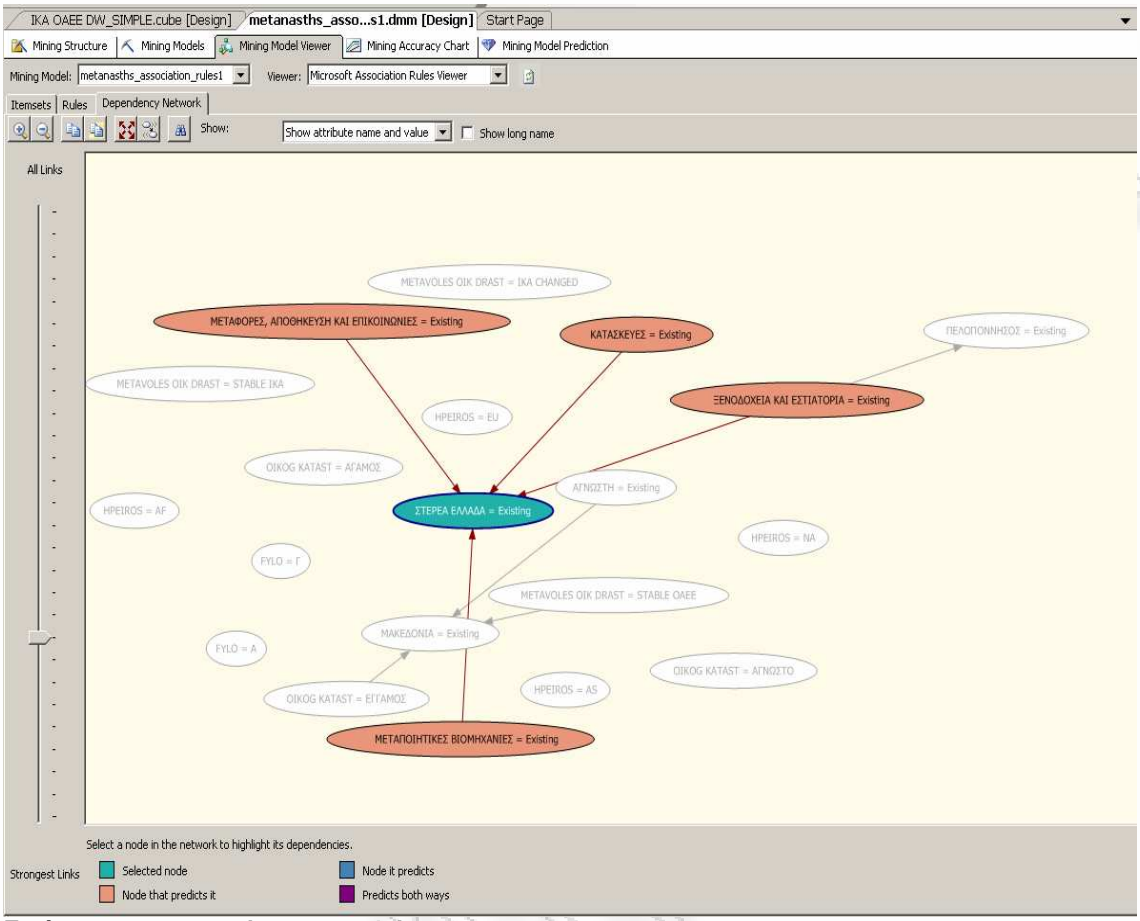
Εικόνα 5-17. Itemsets

Όπως παρατηρούμε στην παραπάνω εικόνα, στην καρτέλα Itemsets απεικονίζεται η υποστήριξη (ο αριθμός των συναλλαγών - εγγραφών που το συγκεκριμένο στοιχειοσύνολο εμφανίζεται), το μέγεθος (ο αριθμός των στοιχείων που αποτελούν το στοιχειοσύνολο) καθώς και η μορφή του στοιχειοσυνόλου με τα πραγματικά του αντικείμενα. Δηλαδή για παράδειγμα, σε 50836 εγγραφές εμφανίζεται το στοιχειοσύνολο {FYLO=A, HPEIROS=EU} μεγέθους 2.



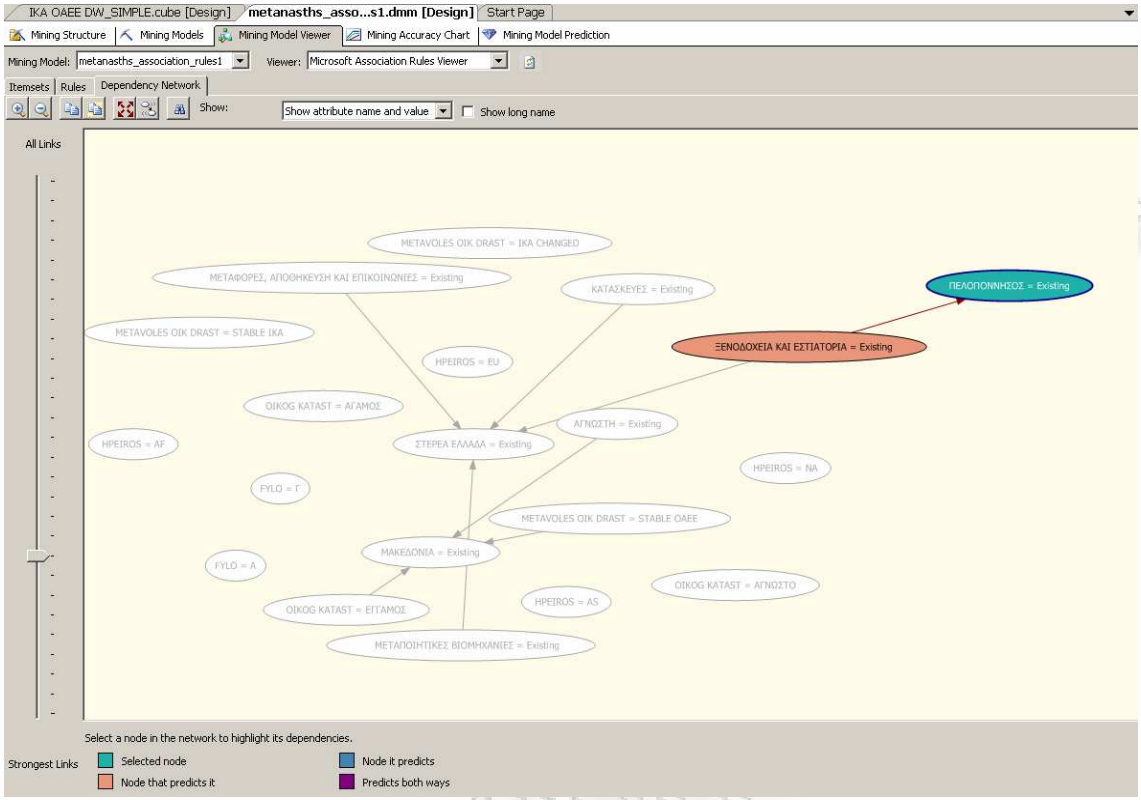
Εικόνα 5-18.Rules

Όπως παρατηρούμε στην παραπάνω Εικόνα, η καρτέλα Rules εμφανίζει τους κανόνες συσχέτισης που προκύπτουν από την ανάλυση των στοιχειοσυνόλων της προηγούμενης καρτέλας, με το ποσοστό εμπιστοσύνης (probability) και το επίπεδο σημαντικότητας (importance) που αντιστοιχούν σε κάθε κανόνα. Επομένως για παράδειγμα, στο 88,7% (ποσοστό εμπιστοσύνης) των στοιχειοσυνόλων που περιλαμβάνουν τα γνωρίσματα {ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ=Existing, HPEIROS=AS} υπάρχει και το γνώρισμα {ΣΤΕΡΕΑ ΕΛΛΑΔΑ= Existing}. Δηλαδή, εάν κάποιος μετανάστης απασχολείται στις ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ (είναι δηλαδή ασφαλισμένος του ΙΚΑ αφού έχει οικονομική δραστηριότητα) και προέρχεται από χώρα της Ασίας είναι πολύ πιθανό να διαμένει και να απασχολείται σε κάποιο νομό της Στερεάς Ελλάδας. Παρατηρώντας πιο προσεκτικά τους εξαγόμενους κανόνες συσχέτισης, θα μπορούσαμε να πούμε ότι επαληθεύουν αποτελέσματα της OLAP ανάλυσης με κάποιο ποσοστό υποστήριξης και κάποιο επίπεδο εμπιστοσύνης.



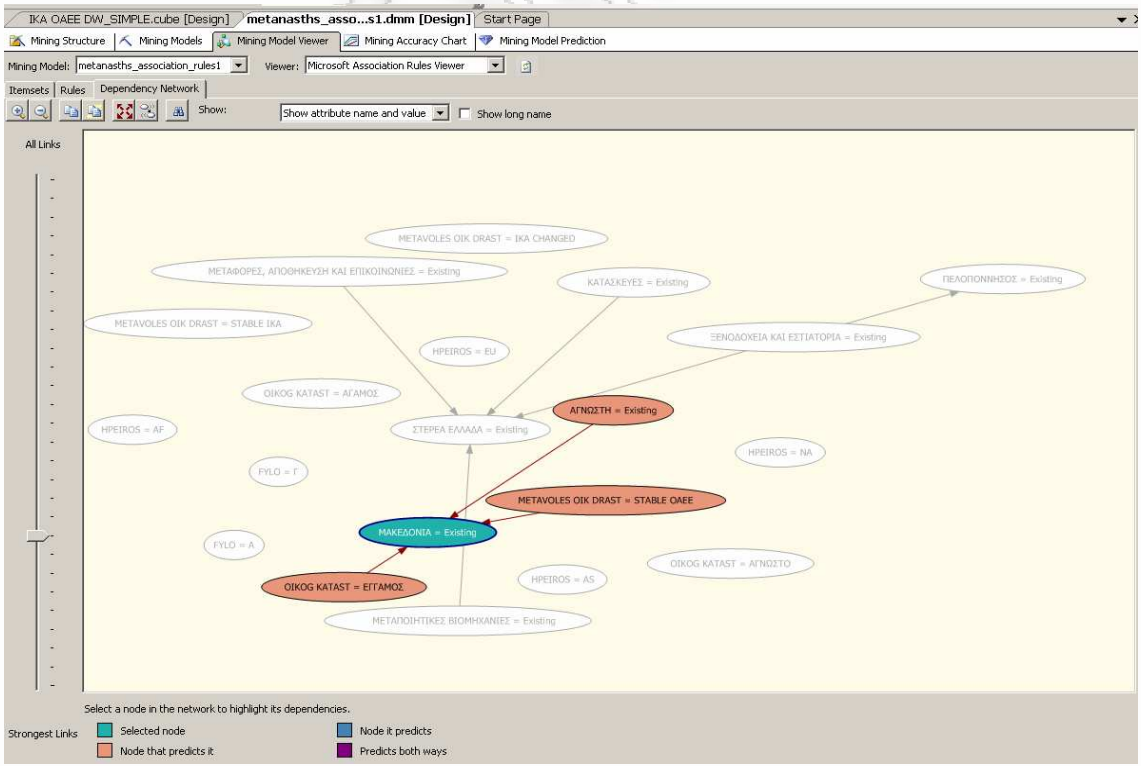
Εικόνα 5-19α. Dependency Network

Με την καρτέλα Dependency Network μπορούμε να διερευνήσουμε την αλληλεπίδραση μεταξύ διαφορετικών αντικειμένων του μοντέλου. Επομένως, όπως παρατηρούμε στην Εικόνα 5-19α όταν κάποιος μετανάστης απασχολείται σε κάποια από τις οικονομικές δραστηριότητες ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ, ΜΕΤΑΦΟΡΕΣ, ΑΠΟΘΗΚΕΥΣΗ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΕΣ, ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ και ΚΑΤΑΣΚΕΥΕΣ, δηλαδή είναι ασφαλισμένος του ΙΚΑ, είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Στερεάς Ελλάδας.



Εικόνα 5-19β. Dependency Network

Όπως παρατηρούμε στην Εικόνα 5-19β όταν κάποιος μετανάστης απασχολείται στην οικονομική δραστηριότητα ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ, δηλαδή είναι ασφαλισμένος του ΙΚΑ, είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Πελοποννήσου.



Εικόνα 5-19γ. Dependency Network

Όπως παρατηρούμε στην Εικόνα 5-19γ όταν κάποιος μετανάστης είναι ασφαλισμένος του ΟΑΕΕ και είναι έγγαμος, είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Μακεδονίας.

5.2.6 Εύρεση Κανόνων Συσχετίσεων Μεταξύ Δημογραφικών Δεδομένων Ασφαλισμένων Μεταναστών (2η περίπτωση)

Το μοντέλο εξόρυξης γνώσης που έχει υλοποιηθεί είναι το metanasths_ association_rules2.dmm, το οποίο έχει ως στόχο την εύρεση συχνών συνόλων αντικειμένων με τα δημογραφικά δεδομένα των ασφαλισμένων μεταναστών στους ασφαλιστικούς οργανισμούς ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ, τους πιθανούς κανόνες συσχέτισης των αντικειμένων, το ποσοστό υποστήριξης και το επίπεδο εμπιστοσύνης. Η υλοποίηση του μοντέλου πραγματοποιήθηκε με χρήση του αλγόριθμου Κανόνων Συσχετίσεων Association Rules του Microsoft SQL Server 2005.

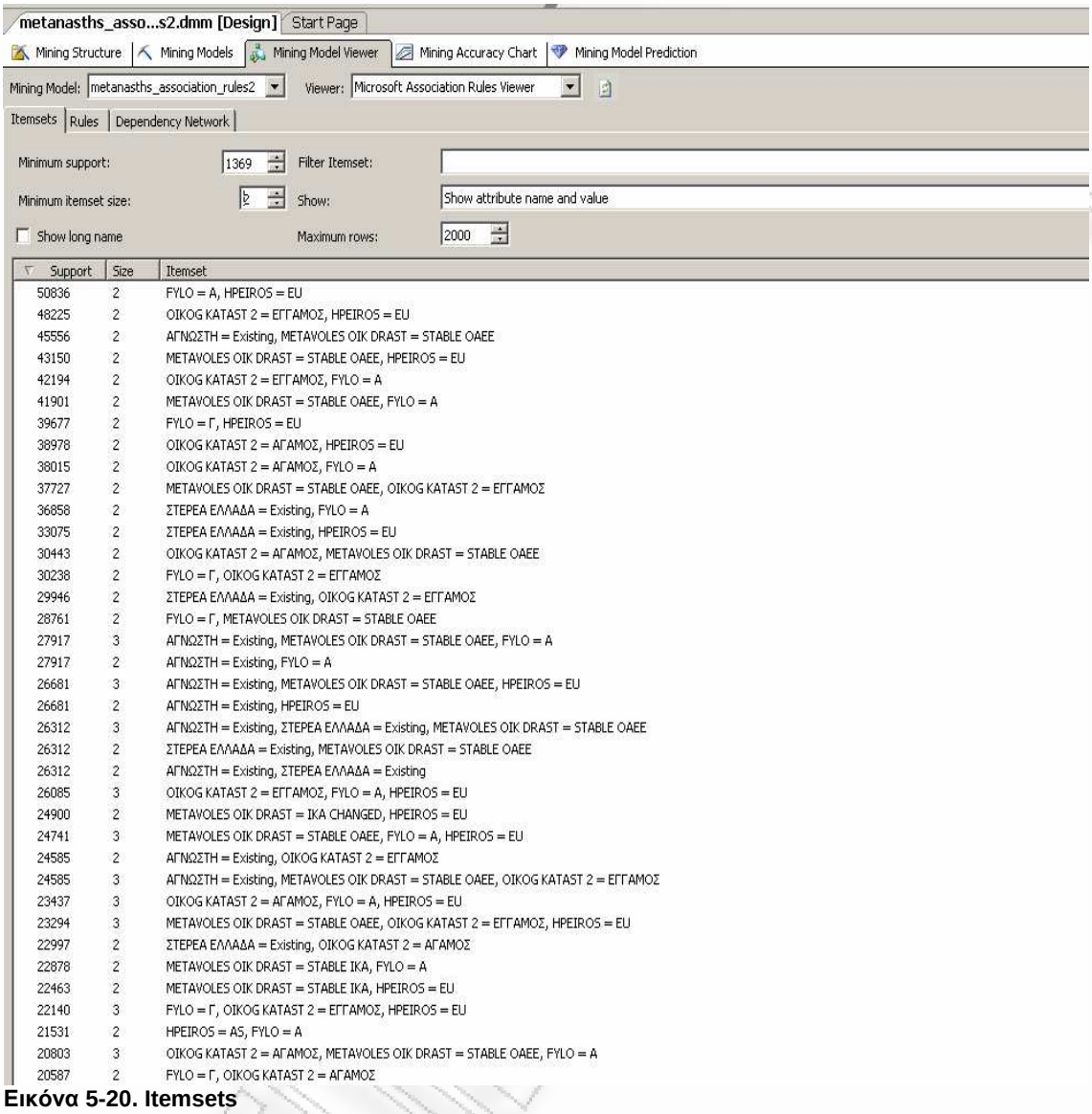
Στον παρακάτω πίνακα δίνονται τα γνωρίσματα που χρησιμοποιήθηκαν στη δημιουργία της δομής του μοντέλου και τα οποία αντλήθηκαν από τους πίνακες του κύβου ΙΚΑ_ΟΑΕΕ_DW_ADVANCED.cube.

Πίνακας 5-8. Γνωρίσματα του μοντέλου εξόρυξης γνώσης metanasths_ association_rules2.dmm

Πίνακες	Στήλες	Κλειδί	Γνωρίσματα Εισόδου	Γνωρίσματα Πρόβλεψης
METANASTHS_DM	METANASTHS_DM _ ADVANCED	X		
	FYLO		X	
	HPEIROS		X	
	METAVOLES_OIK_ DRAST		X	
	OIKOG_KATAST_2		X	
OIKONOM_DRAST_ DM	DESC_OIKONOM_ DRAST	X	X	
PER_KATOIKIAS	DIAMERISMA	X		X

Επίσης, στον κύβο πραγματοποιήθηκε ίδιος τεμαχισμός (slice) και ορίστηκαν τα ίδια φίλτρα της προηγούμενης παραγράφου. Επίσης, στις παραμέτρους του αλγορίθμου MINIMUM_PROBABILITY (ελάχιστη πιθανότητα) και MINIMUM_SUPPORT (ελάχιστη υποστήριξη) τέθηκαν πάλι οι τιμές 0,1 και 0,01 αντίστοιχα.

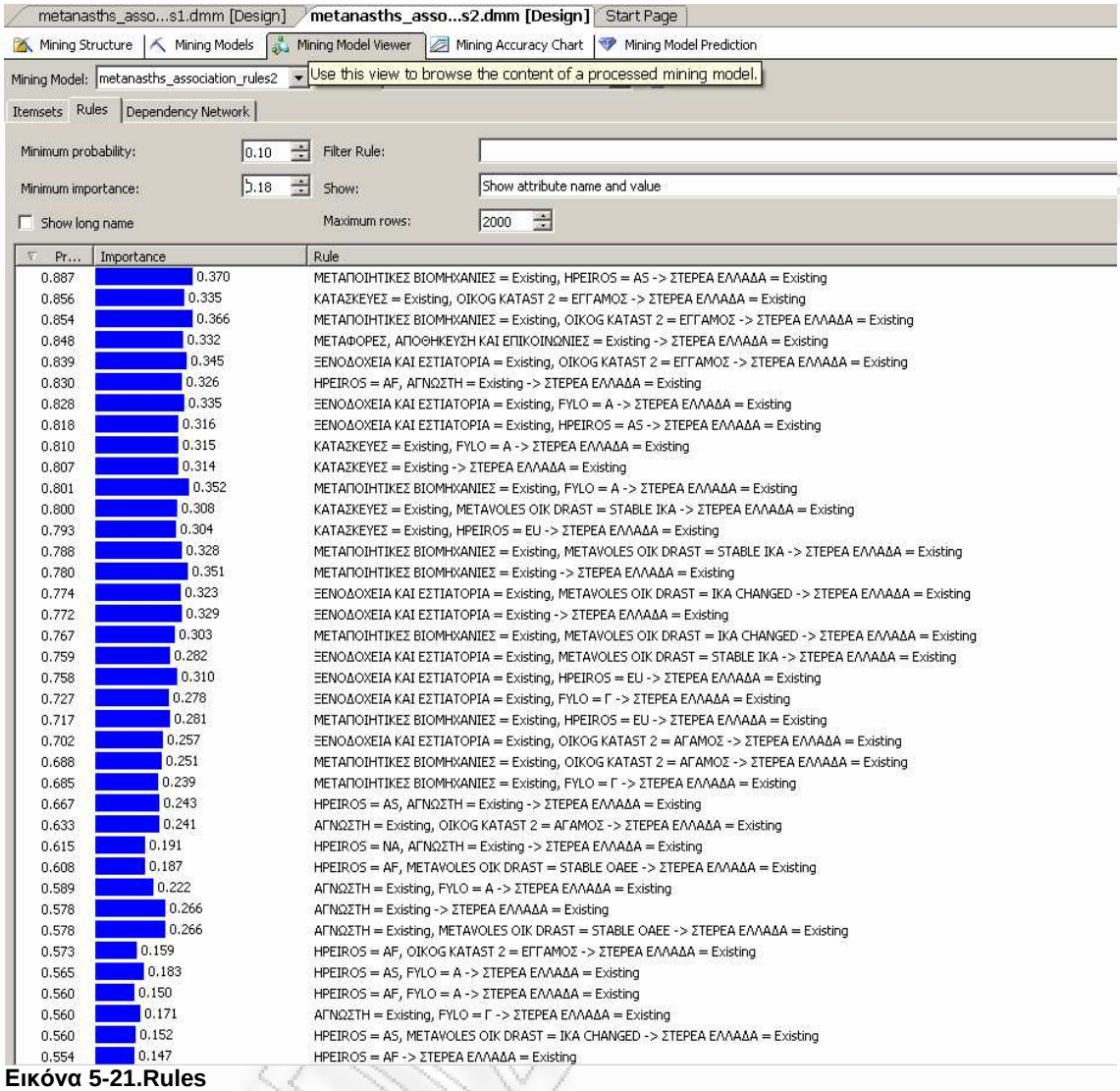
Στη συνέχεια, ακολουθεί η διερεύνηση του μοντέλου εξόρυξης γνώσης μελετώντας τα στιγμιότυπα των καρτελών Itemsets (στοιχειοσύνολα), Rules (κανόνες) και Dependency Network (διάγραμμα συσχετίσεων).



Support	Size	Itemset
50836	2	FYLO = A, HPEIROS = EU
48225	2	OIKOG KATAST 2 = ΕΓΓΑΜΟΣ, HPEIROS = EU
45556	2	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE
43150	2	METAVOLES OIK DRAST = STABLE OAEE, HPEIROS = EU
42194	2	OIKOG KATAST 2 = ΕΓΓΑΜΟΣ, FYLO = A
41901	2	METAVOLES OIK DRAST = STABLE OAEE, FYLO = A
39677	2	FYLO = Γ, HPEIROS = EU
38978	2	OIKOG KATAST 2 = ΑΓΑΜΟΣ, HPEIROS = EU
38015	2	OIKOG KATAST 2 = ΑΓΑΜΟΣ, FYLO = A
37727	2	METAVOLES OIK DRAST = STABLE OAEE, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ
36858	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, FYLO = A
33075	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, HPEIROS = EU
30443	2	OIKOG KATAST 2 = ΑΓΑΜΟΣ, METAVOLES OIK DRAST = STABLE OAEE
30238	2	FYLO = Γ, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ
29946	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ
28761	2	FYLO = Γ, METAVOLES OIK DRAST = STABLE OAEE
27917	3	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE, FYLO = A
27917	2	ΑΓΝΩΣΤΗ = Existing, FYLO = A
26681	3	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE, HPEIROS = EU
26681	2	ΑΓΝΩΣΤΗ = Existing, HPEIROS = EU
26312	3	ΑΓΝΩΣΤΗ = Existing, ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, METAVOLES OIK DRAST = STABLE OAEE
26312	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, METAVOLES OIK DRAST = STABLE OAEE
26312	2	ΑΓΝΩΣΤΗ = Existing, ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing
26085	3	OIKOG KATAST 2 = ΕΓΓΑΜΟΣ, FYLO = A, HPEIROS = EU
24900	2	METAVOLES OIK DRAST = IKA CHANGED, HPEIROS = EU
24741	3	METAVOLES OIK DRAST = STABLE OAEE, FYLO = A, HPEIROS = EU
24585	2	ΑΓΝΩΣΤΗ = Existing, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ
24585	3	ΑΓΝΩΣΤΗ = Existing, METAVOLES OIK DRAST = STABLE OAEE, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ
23437	3	OIKOG KATAST 2 = ΑΓΑΜΟΣ, FYLO = A, HPEIROS = EU
23294	3	METAVOLES OIK DRAST = STABLE OAEE, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ, HPEIROS = EU
22997	2	ΣΤΕΡΕΑ ΕΛΛΑΔΑ = Existing, OIKOG KATAST 2 = ΑΓΑΜΟΣ
22878	2	METAVOLES OIK DRAST = STABLE IKA, FYLO = A
22463	2	METAVOLES OIK DRAST = STABLE IKA, HPEIROS = EU
22140	3	FYLO = Γ, OIKOG KATAST 2 = ΕΓΓΑΜΟΣ, HPEIROS = EU
21531	2	HPEIROS = AS, FYLO = A
20803	3	OIKOG KATAST 2 = ΑΓΑΜΟΣ, METAVOLES OIK DRAST = STABLE OAEE, FYLO = A
20587	2	FYLO = Γ, OIKOG KATAST 2 = ΑΓΑΜΟΣ

Εικόνα 5-20. Itemsets

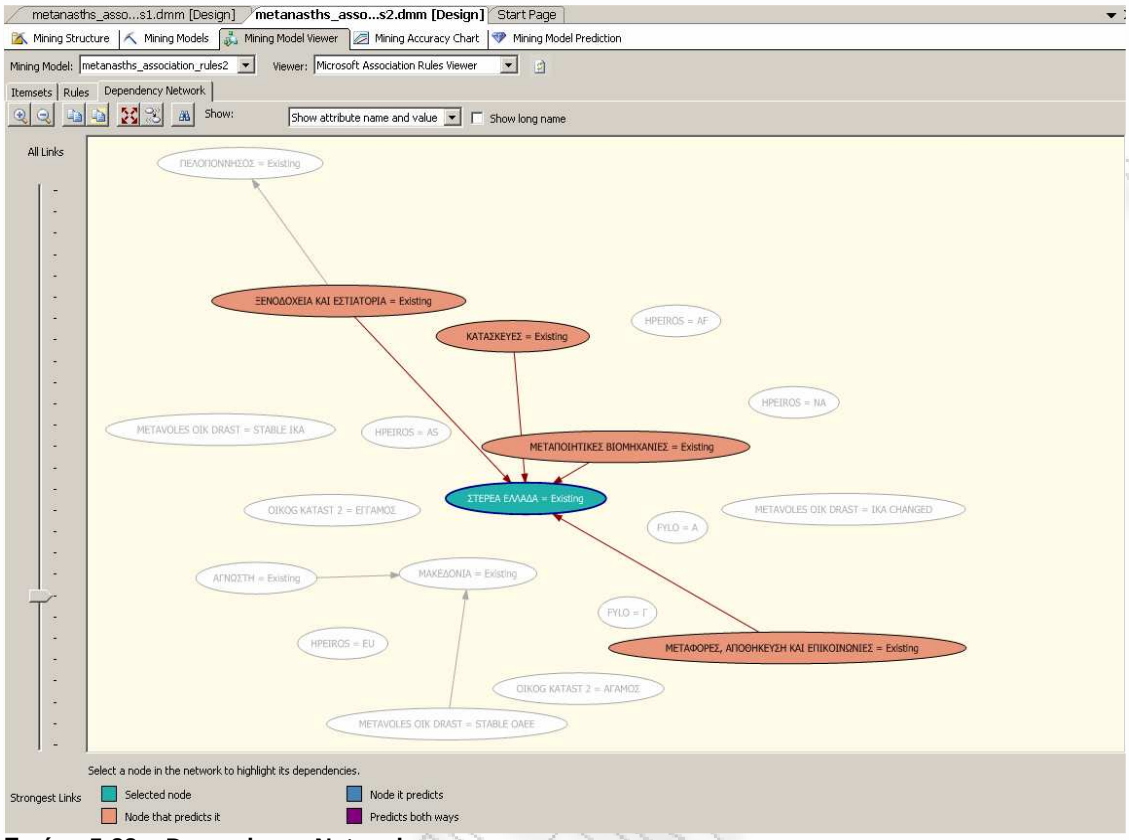
Όπως παρατηρούμε στην παραπάνω εικόνα, τα αποτελέσματα είναι περίπου ίδια με αυτά του μοντέλου metanasths_ association_rules1.dmm παρόλο που χρησιμοποιήθηκε το γνώρισμα OIKOG_KATAST_2 αντί του OIKOG_KATAST, το οποίο περιέχει στοιχεία οικογενειακής κατάστασης και για τους ασφαλισμένους του ΙΚΑ. Δηλαδή και πάλι σε 50836 εγγραφές εμφανίζεται το στοιχειοσύνολο {FYLO=A, HPEIROS=EU} μεγέθους 2, ενώ τα στοιχειοσύνολα που ακολουθούν διαφέρουν ελαφρώς από αυτά του μοντέλου της προηγούμενης παραγράφου.



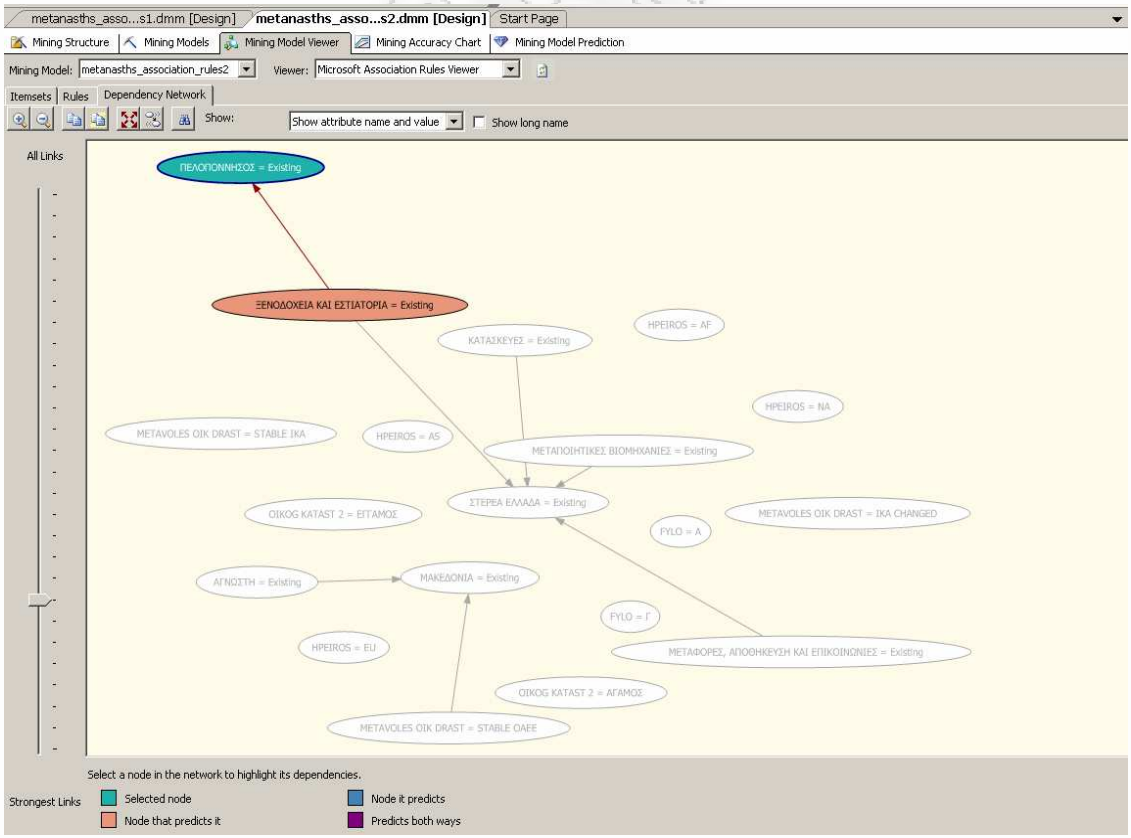
Εικόνα 5-21.Rules

Όπως παρατηρούμε και η καρτέλα Rules δεν εμφανίζει σημαντικές διαφορές με την αντίστοιχη της προηγούμενης παραγράφου.

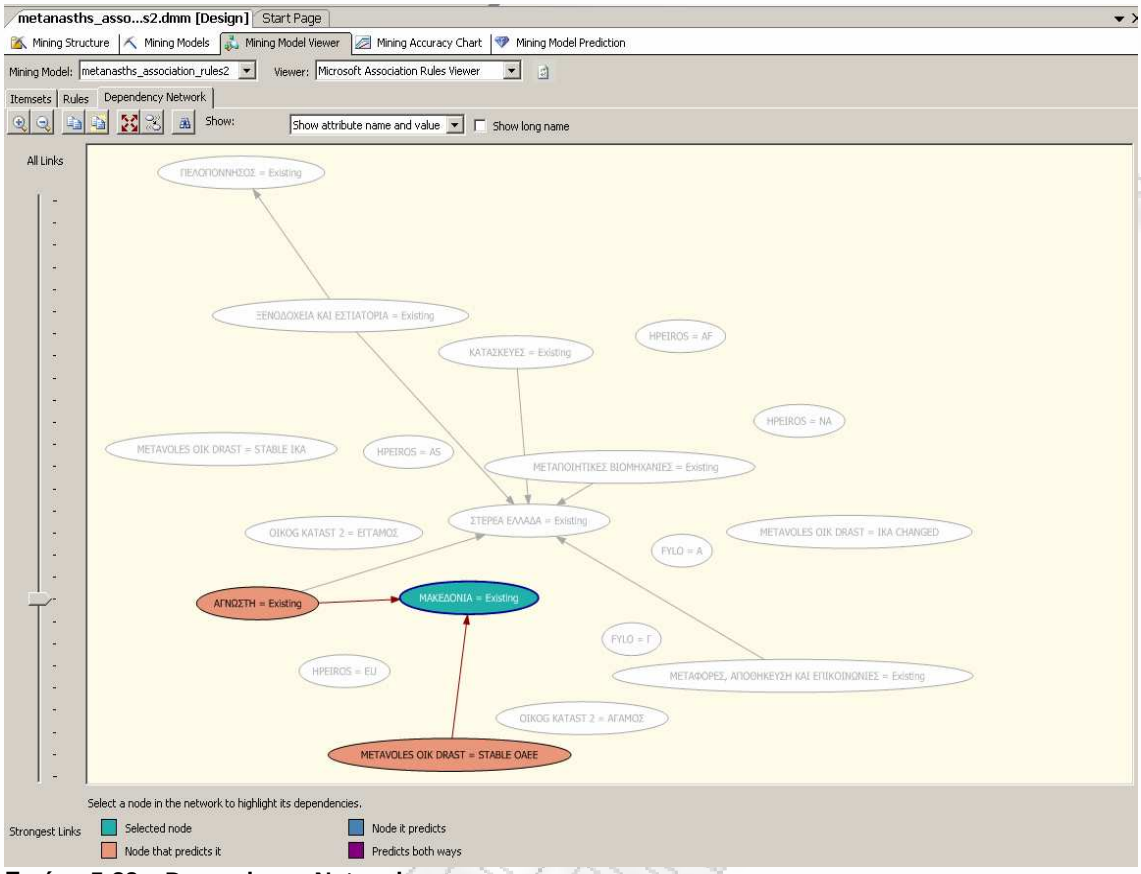
Ολοκληρώνοντας τη διερεύνηση του μοντέλου με τη μελέτη των εξαγόμενων αποτελεσμάτων της καρτέλας Dependency Network όπως αυτά απεικονίζονται στις παρακάτω Εικόνες 5-22α, 5-22β και 5-22γ διαπιστώνουμε ότι το μοντέλο metanasths_association_ rules2.dmm δεν διαφέρει από το αντίστοιχο της προηγούμενης παραγράφου. Δηλαδή και πάλι, ο μετανάστης ασφαλισμένος του ΙΚΑ που απασχολείται σε κάποια από τις οικονομικές δραστηριότητες ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ, ΜΕΤΑΦΟΡΕΣ, ΑΠΟΘΗΚΕΥΣΗ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΕΣ, ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ και ΚΑΤΑΣΚΕΥΕΣ είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Στερεάς Ελλάδας. Επίσης, ο ασφαλισμένος μετανάστης του ΙΚΑ που απασχολείται στην οικονομική δραστηριότητα ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Πελοποννήσου. Τέλος, όταν κάποιος μετανάστης είναι ασφαλισμένος του ΟΑΕΕ είναι πολύ πιθανό να διαμένει και δραστηριοποιείται σε κάποιο νομό της Μακεδονίας. Επομένως, το εξαγόμενο συμπέρασμα είναι ότι η οικογενειακή κατάσταση του μετανάστη δεν αποτελεί τόσο σημαντικό γνώρισμα προσδιορισμού του τόπου κατοικίας και εργασίας του κάθε μετανάστη όσο η οικονομική του δραστηριότητα, η μεταβολή αυτής ή η ήπειρος προέλευσής του.



Εικόνα 5-22α. Dependency Network



Εικόνα 5-22β. Dependency Network



Εικόνα 5-22γ. Dependency Network

6. Παρατηρήσεις – Συμπεράσματα

6.1 Παρατηρήσεις

Σκοπός της συγκεκριμένης παραγράφου είναι η παράθεση ορισμένων παρατηρήσεων που σχετίζονται με την εφαρμογή τεχνικών εξόρυξης γνώσης σε πολυδιάστατα μοντέλα δεδομένων – κύβους που παρουσιάστηκαν στο Κεφάλαιο 5. Οι παρατηρήσεις αυτές στοχεύουν στην πληρέστερη κατανόηση των διαδικασιών που ακολουθήθηκαν για την εξαγωγή των επιθυμητών αποτελεσμάτων.

Όπως είδαμε στο σχετικό κεφάλαιο, η υλοποίηση των μοντέλων εξόρυξης γνώσης με την εφαρμογή διαδικασιών Κατηγοριοποίησης (Classification), Συσταδοποίησης (Clustering) και Εύρεσης Κανόνων Συσχετίσεων (Association Rules) πραγματοποιήθηκε με τη χρησιμοποίηση του εργαλείου Analysis Services του Microsoft SQL Server 2005. Το εργαλείο διαθέτει συνολικά εννέα αλγόριθμους (Association Rules, Decision Trees, Naïve Bayes, Neural Nets, Clustering, Sequence Clustering, Time Series) οι οποίοι μπορούν να εφαρμοστούν στα δεδομένα σχεσιακών βάσεων ή σε κύβους.

Για τις ανάγκες της παρούσας εργασίας επιλέχθηκε η εφαρμογή των αλγόριθμων Decision Trees, Clustering και Association Rules σε κύβους και όχι στα σχεσιακά δεδομένα της Αποθήκης, γεγονός που δημιούργησε ορισμένες δυσκολίες κατά την ανάλυση οι οποίες πρέπει να επισημανθούν καθώς αποτελούν αδυναμίες του εργαλείου εξόρυξης γνώσης Analysis Services. Συνοπτικά, μπορούμε να αναφέρουμε τα εξής:

- Το εργαλείο δεν υποστηρίζει κύβους με σχήμα χιονονιφάδας (snowflake schema).
- Το εργαλείο δεν υποστηρίζει τις λειτουργίες του ελέγχου της ακρίβειας των μοντέλων εξόρυξης γνώσης (mining accuracy chart) και της δημιουργίας προβλέψεων από τα μοντέλα εξόρυξης γνώσης (mining model prediction), όταν τα μοντέλα αυτά δημιουργούνται πάνω σε κύβους.
- Διαστάσεις που περιέχουν πεδία με κενές τιμές δεν μπορούν να χρησιμοποιηθούν στα μοντέλα εξόρυξης γνώσης. Πρέπει να συμπληρωθούν οι ελλιπείς τιμές ακόμη και με τη χρήση μιας γενικής σταθεράς (π.χ. ΑΓΝΩΣΤΟ), όπως και έγινε στα σχετικά παραδείγματα.
- Διαστάσεις που δημιουργούνται στους κύβους με σκοπό την διακριτοποίηση μετρήσιμων μεγεθών (π.χ. Cube dimensions PALAIOTHTA, HLIKIAKES OMADES), δεν μπορούν να χρησιμοποιηθούν ως εμφωλευμένοι πίνακες στα μοντέλα εξόρυξης γνώσης.
- Όταν υπάρχουν γνωρίσματα – πεδία διαστάσεων με περισσότερες από 100 καταστάσεις (π.χ. οι χώρες καταγωγής που περιλαμβάνονται στην Αποθήκη είναι 241) τότε οι αλγόριθμοι επιλέγουν τυχαία τις πρώτες 100 καταστάσεις τόσο στα εξαγόμενα αποτελέσματα όσο και στον τεμαχισμό (slice) του κύβου. Αυτός είναι και ο λόγος που στα σχετικά παραδείγματα χρησιμοποιήσαμε ως γνώρισμα την περιγραφή των ηπείρων αντί της περιγραφής των χωρών (6 αντί 241 καταστάσεων).

Επομένως, με βάση τις παραπάνω παρατηρήσεις και με σχετικές συμβάσεις και παραδοχές που ορίστηκαν, δημιουργήθηκαν πιο ευέλικτα σχήματα για τους κύβους που χρησιμοποιήθηκαν στη εξόρυξη γνώσης από αυτά της OLAP ανάλυσης. Αυτό είχε ως αποτέλεσμα να ξεπεραστούν τελικά οι σχετικές δυσκολίες και η εκτέλεση των αλγόριθμων να οδηγήσει στην εξαγωγή χρήσιμης πληροφορίας από τα δημογραφικά δεδομένα ασφαλισμένων μεταναστών στο ελλαδικό χώρο.

Συνοψίζοντας, οι κύβοι αν και παρέχουν τη δυνατότητα επιλογής ιεραρχιών ως φίλτρα των διαστάσεων των δεδομένων, πρακτικά διαθέτουν μικρή ευελιξία σε σύγκριση με τα σχεσιακά δεδομένα όταν επιλέγονται ως πηγή δεδομένων για τη δημιουργία μοντέλων εξόρυξης γνώσης. Αυτός μάλλον είναι και ο λόγος που στη σχετική βιβλιογραφία όλα τα παραδείγματα δημιουργίας μοντέλων εξόρυξης γνώσης από δεδομένα υλοποιούνται πάνω σε σχεσιακούς πίνακες βάσεων δεδομένων ή ακόμα καλύτερα σε κατάλληλα διαμορφωμένες όψεις (views) σχεσιακών πινάκων ώστε να εξασφαλίζεται η πληρότητα και η αρτιότητα των στοιχείων και άρα η ποιότητα των εξαγόμενων αποτελεσμάτων.

6.2 Συμπεράσματα

Τα τελευταία χρόνια, η εξέλιξη των υπολογιστικών συστημάτων και των τεχνολογιών αποθήκευσης δεδομένων σε συνδυασμό με τη διάδοση των συστημάτων διαχείρισης βάσεων δεδομένων οδήγησαν στη συγκέντρωση και αποθήκευση τεράστιου όγκου πληροφορίας κάθε είδους. Η σύγχρονη κοινωνία, αναφέρεται συχνά ως κοινωνία της πληροφορίας καθώς για κάθε είδος πληροφορίας που παράγεται υπάρχει η δυνατότητα καταγραφής του και άρα αποθήκευσής του, με αποτέλεσμα να διογκώνεται το μέγεθος των αντίστοιχων βάσεων δεδομένων. Ωστόσο, η συνολική διαχείριση και αξιοποίηση του τεράστιου όγκου της διαθέσιμης πληροφορίας αποτελεί ένα σημαντικό και δύσκολο έργο το οποίο μπορεί να αντιμετωπιστεί αποτελεσματικά μόνο με τη χρήση σύγχρονων τεχνολογιών πληροφορικής για την ανάπτυξη προηγμένων συστημάτων διαχείρισης βάσεων δεδομένων και υποστήριξης λήψης αποφάσεων.

Το έναυσμα για την εκπόνηση της εργασίας προήλθε από τη διακήρυξη ανοιχτού διαγωνισμού του Ινστιτούτου Μεταναστευτικής Πολιτικής (Ι.ΜΕ.ΠΟ) για Πληροφοριακό Σύστημα Διαχείρισης και Αξιοποίησης Μεταναστευτικών Δεδομένων που θα εξυπηρετεί τη συλλογή, διασταύρωση, διαχείριση και επιχειρησιακή αξιοποίηση δεδομένων των φορέων Υπουργείο Εσωτερικών, Αποκέντρωσης & Ηλεκτρονικής Διακυβέρνησης, ΙΚΑ-ΕΤΑΜ, ΟΑΕΕ/ΤΕΒΕ και ΟΓΑ, με στόχο την εξαγωγή αναλυτικών συμπερασμάτων για τη χάραξη εμπεριστατωμένης μεταναστευτικής πολιτικής.

Στα πλαίσια υλοποίησης της παρούσας εργασίας παρουσιάζεται η αξιοποίηση ώριμων τεχνολογικών λύσεων διαχείρισης και ανάλυσης της πληροφορίας που προέρχεται από σχεσιακές βάσεις δεδομένων ή άλλες πηγές πρωτογενών δεδομένων (flat αρχεία κλπ.). Ουσιαστικά, υλοποιήθηκε σε μικρή κλίμακα ένα πολυδιάστατο σύστημα συλλογής, αποθήκευσης, επεξεργασίας και αξιοποίησης μεταναστευτικών δεδομένων των ενεργών αλλοδαπών ασφαλισμένων των Ασφαλιστικών Οργανισμών ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ/ΤΕΒΕ με την εφαρμογή σύγχρονων τεχνολογιών όπως είναι οι αποθήκες δεδομένων (data warehouses), οι λειτουργίες αναλυτικής επεξεργασίας δεδομένων (OLAP) και οι τεχνικές εξόρυξης γνώσης (data mining techniques) ώστε να καταφανούν και να σχολιαστούν οι ιδιαιτερότητες και οι δυσκολίες που κρύβει η μετατροπή δεδομένων από διάφορες πηγές σε πληροφορία και άρα σε γνώση, με κεντρικό σημείο αναφοράς το μετανάστη.

Το πρακτικό μέρος της εργασίας αφορά την υλοποίηση Αποθήκης Μεταναστευτικών Δεδομένων που να συγκεντρώνει τα ετερογενή δημογραφικά δεδομένα των ασφαλισμένων μεταναστών των Ασφαλιστικών Οργανισμών ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ/ΤΕΒΕ, την ανάλυση τύπου OLAP και την εφαρμογή τεχνικών εξόρυξης γνώσης στα δεδομένα αυτά με τη χρήση του Συστήματος Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) του Microsoft SQL Server 2005 και του λογισμικού Analysis Services.

Το κύριο πρόβλημα που αντιμετωπίσαμε κατά το λογικό σχεδιασμό της Αποθήκης Δεδομένων ήταν η αδυναμία πρόσβασης στα πρωτογενή δεδομένα. Η αρχιτεκτονική και το λογικό μοντέλο που τελικά ακολουθήθηκε κατά τον φυσικό σχεδιασμό της Αποθήκης Δεδομένων υπαγορεύτηκε από τη δομή και το περιεχόμενο των βάσεων δεδομένων του ΙΚΑ-ΕΤΑΜ και του ΟΑΕΕ ώστε να επιτευχθεί ο συγκερασμός των δεδομένων των δύο ετερογενών πηγών και να διατηρηθεί αυτούσια η διαθέσιμη πληροφορία που δεν μπορούσε να συνενωθεί. Αρκετά από τα δεδομένα του ΙΚΑ παράχθηκαν βάσει επίσημων δημοσιευμένων εξαμηνιαίων στατιστικών στοιχείων απασχόλησης των αλλοδαπών ασφαλισμένων και στοιχείων της Εθνικής Στατιστικής Υπηρεσίας, και ουσιαστικά αυτά καθόρισαν και την υλοποίηση της σχεσιακής βάσης δεδομένων του ΙΚΑ. Παράλληλα, για τις ανάγκες της εργασίας, που σκοπός της είναι η αξιοποίηση σύγχρονων τεχνολογιών πληροφορικής για τη μετατροπή των δεδομένων που συγκεντρώνονται από διάφορες πηγές, σε πληροφορία και άρα σε γνώση, μας παραχωρήθηκαν δημογραφικά και οικονομικά δεδομένα που αφορούν ασφαλισμένους αλλοδαπούς του ΟΑΕΕ. Τα δεδομένα αυτά είχαν συγκεντρωθεί σε σχεσιακή βάση που υλοποιήθηκε επίσης στο ΣΔΒΔ του Microsoft SQL Server 2005, στα πλαίσια εκπόνησης άλλης μεταπτυχιακής διατριβής.

Στη συνέχεια, συγκεντρώθηκαν στην Αποθήκη Δεδομένων τα ετερογενή δημογραφικά και οικονομικά δεδομένα από τις δύο βάσεις δεδομένων (ΙΚΑ-ΕΤΑΜ και ΟΑΕΕ/ΤΕΒΕ) με κατάλληλες ETL διαδικασίες, ώστε να είναι τελικά δυνατή η εφαρμογή σε αυτά λειτουργιών αναλυτικής επεξεργασίας και τεχνικών εξόρυξης γνώσης. Ωστόσο, παρόλο που ορίστηκαν

συγκεκριμένες παραδοχές και συμβάσεις κατά τη φόρτωση των δεδομένων από τις βάσεις στην Αποθήκη, ώστε να εξασφαλιστεί η συνέπεια των δεδομένων, κάποιες ιδιαιτερότητες δεν ήταν εφικτό να ξεπεραστούν. Όπως ήταν αναμενόμενο, οι ιδιαιτερότητες αυτές καταφάνηκαν στις παραγράφους που παρουσιάστηκαν ενδεικτικά παραδείγματα εφαρμογής λειτουργιών OLAP καθώς και δημιουργίας μοντέλων εξόρυξης γνώσης σε πολυδιάστατα μοντέλα δεδομένων – κύβους και καταγράφηκαν τα σχετικά συμπεράσματα.

Ουσιαστικά λόγω της έλλειψης πρωτογενών δεδομένων αλλά και της ανομοιογένειας των διαθέσιμων δεδομένων, σε κάθε στάδιο της ανάλυσης είναι εμφανής ο διαχωρισμός το συνόλου των αλλοδαπών ασφαλισμένων που μελετάμε σε δύο πληθυσμούς. Ο ένας πληθυσμός αποτελείται από ασφαλισμένους που απασχολούνται με εξαρτημένη εργασία (ασφαλισμένοι ΙΚΑ), για τους οποίους γνωρίζουμε τα δημογραφικά τους στοιχεία (πλην της οικογενειακής τους κατάστασης) και τις μεταβολές της οικονομικής τους δραστηριότητας για μια σειρά ετών. Ο δεύτερος πληθυσμός αποτελείται από ελεύθερους επαγγελματίες (ασφαλισμένοι ΟΑΕΕ/ΤΕΒΕ), για τους οποίους γνωρίζουμε τα δημογραφικά και οικονομικά τους στοιχεία (καταβληθείσες εισφορές) για μια περίοδο ετών που επικαλύπτεται με την αντίστοιχη του ΙΚΑ, όμως δεν γνωρίζουμε καθόλου το αντικείμενο της οικονομικής τους δραστηριότητας.

Επομένως, στα πλαίσια βελτίωσης του παρόντος παραδοτέου, προτείνεται η εισαγωγή νέων, πληρέστερων πρωτογενών δεδομένων ώστε να επιτευχθεί μια πιο αποτελεσματική συνένωση των μητρώων αλλοδαπών ασφαλισμένων των δύο ασφαλιστικών οργανισμών που θα οδηγήσει στην εξαγωγή πιο εμπεριστατωμένων συμπερασμάτων σχετικά με το συνολικό προφίλ του μετανάστη που διαμένει και δραστηριοποιείται στον ελλαδικό χώρο, με βάση πάντα τα δημογραφικά του στοιχεία, την εργασιακή του κινητικότητα και την ασφαλιστική του ιστορία.

7. Βιβλιογραφικές Αναφορές

1. Berry, M.J.A., Linoff, G.S.: Data mining techniques : for marketing, sales, and customer relationship management. Wiley Publishing, Inc. (2004)
2. Dasu, T., Johnson, T.: Exploratory Data Mining and Data Cleaning. John Wiley & Sons. (2003)
3. Dunham, M.H.: Data Mining. Επιμέλεια έκδοσης: Β. Βερύκιος, Γ. Θεοδωρίδης. Εκδόσεις Νέων Τεχνολογιών. (2004)
4. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.: The Knowledge Discovery in Databases – kdd process for extracting useful knowledge from volumes of data. Journal of the Acm, (November 1996).
5. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. (1996)
6. Fayyad, U., Grinstein, G., Wierse, A.: Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann. (2001)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed. (2006)
8. Hand, D. J., Mannila, H., Smyth, P.: Principles of Data Mining, MIT Press. (2001)
9. Imhoff, C., Galemmo, N., Geiger, J. G.: Mastering Data Warehouse Design: Relational and Dimensional Techniques. Wiley Publishing, Inc. (2003)
10. Inmon, W.H.: The data warehouse and data mining. Communications of the ACM. (November 1996)
11. Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons, Inc. (2002)
12. Jacobson, Reed, Misner, Stacia: Micropsoft SQL SERVER 2005 ANALYSIS SERVICES Step by Step, Microsoft Press. (2006)
13. Kimball, Ralph, Ross, Margy: The Data Warehouse Toolkit. John Wiley & Sons, Inc. (2002)
14. Marakas, George M.: Modern data Warehousing, Mining and Visualization. Pearson Education, Inc. (2003)
15. Nielsen, Paul.: SQL Server™ 2005 Bible. Wiley Publishing, Inc. (2007)
16. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, Wiley, (2005)
17. Tang, ZaoHui, MccLennan, Jamie: Data Mining with SQL Server 2005, Wiley Publishing, Inc. (2005)
18. Χαλκίδη, Μ., Βαζιργιάννης, Μ.: ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ. ΤΥΠΩΘΗΤΩ, ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ, Δεύτερη έκδοση. (2005)
19. http://www.ika.gr
20. http://www.imepo.gr/news-competitions-gr.htm
21. http://www.kimballgroup.com/html/designtips.html
22. http://msdn.microsoft.com/en-us/library
23. http://osyk.gandg.gr/cgi-bin/_kad.pl
24. http://www.statistics.gr
25. http://technet.microsoft.com/en-us/library

8. Παραρτήματα

A. SQL Scripts, MDX Scripts

A.1 SQL Scripts δημιουργίας των πινάκων της Βάσης Δεδομένων ΙΚΑ-ΕΤΑΜ

```
CREATE TABLE [dbo].[ASFALISMENOS](
	[AMA] [nchar](9) COLLATE Greek_CI_AS NOT NULL,
	[SHMERINO_EPONYMO] [nchar](50) COLLATE Greek_CI_AS NULL,
	[KYRIO_ONOMA] [nchar](30) COLLATE Greek_CI_AS NULL,
	[EPONYMO_PATROS] [nchar](50) COLLATE Greek_CI_AS NULL,
	[ONOMA_PATROS] [nchar](30) COLLATE Greek_CI_AS NULL,
	[EPONYMO_MHTROS] [nchar](50) COLLATE Greek_CI_AS NULL,
	[ONOMA_MHTROS] [nchar](30) COLLATE Greek_CI_AS NULL,
	[EPONYMO_SYZYGOU] [nchar](50) COLLATE Greek_CI_AS NULL,
	[ONOMA_SYZYGOU] [nchar](30) COLLATE Greek_CI_AS NULL,
	[SEX] [nchar](1) COLLATE Greek_CI_AS NULL,
	[DATE_BIRTH] [datetime] NULL,
	[NATIONALITY_ID] [nchar](2) COLLATE Greek_CI_AS NOT NULL,
	[ETOS_APOGRAFHS] [int] NULL,
	[ODOS_KATOIKIAS] [nchar](50) COLLATE Greek_CI_AS NULL,
	[ARITHMOS_KATOIKIAS] [nchar](5) COLLATE Greek_CI_AS NULL,
	[TK_KATOIKIAS] [nchar](5) COLLATE Greek_CI_AS NULL,
	[NOMOS_ID] [int] NOT NULL,
	[ARITHMOS_TAFTOTHTAS] [nchar](15) COLLATE Greek_CI_AS NULL,
	[AFM] [nchar](9) COLLATE Greek_CI_AS NULL,
	[ARMODIA_DOY] [nchar](20) COLLATE Greek_CI_AS NULL,
	[OIKOGENEIAKH_KATAST] [nchar](10) COLLATE Greek_CI_AS NULL,
	[PROSTATEVOMENA_MELH] [int] NULL,
 CONSTRAINT [PK_ASFALISMENOS] PRIMARY KEY CLUSTERED
(
	[AMA] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

GO
USE [IKA]
GO
ALTER TABLE [dbo].[ASFALISMENOS]  WITH CHECK ADD  CONSTRAINT
[FK_ASFALISMENOS_NATIONALITY] FOREIGN KEY([NATIONALITY_ID])
REFERENCES [dbo].[NATIONALITY] ([NATIONALITY_ID])
GO
ALTER TABLE [dbo].[ASFALISMENOS]  WITH CHECK ADD  CONSTRAINT
[FK_ASFALISMENOS_NOMOI] FOREIGN KEY([NOMOS_ID])
REFERENCES [dbo].[NOMOI] ([NOMOS_ID])

CREATE TABLE [dbo].[PINAKAS_METAVOLWN](
	[AMA] [nchar](9) COLLATE Greek_CI_AS NOT NULL,
	[EXAMHNO] [nchar](10) COLLATE Greek_CI_AS NOT NULL,
	[ETOS] [int] NOT NULL,
	[OIKONOMIKH_DRASTHRIOTHTA_ID] [int] NOT NULL,
	[TYPOS_APASCHOLHSHS] [nchar](100) COLLATE Greek_CI_AS NULL,
	[EIDIKOTHTA] [nchar](50) COLLATE Greek_CI_AS NULL,
	[HMERES_ASFALISHS] [smallint] NULL,
	[HMEROMISTHIO] [money] NULL,
	[EISFORES_ASFALISMENOU] [money] NULL,
	[APODOCHES] [money] NULL,
```

```
 CONSTRAINT [PK_PINAKAS_METAVOLWN] PRIMARY KEY CLUSTERED
(
	[AMA] ASC,
	[EXAMHNO] ASC,
	[ETOS] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

GO
USE [IKA]
GO
ALTER TABLE [dbo].[PINAKAS_METAVOLWN]  WITH CHECK ADD  CONSTRAINT
[FK_PINAKAS_METAVOLWN_ASFALISMENOS] FOREIGN KEY([AMA])
REFERENCES [dbo].[ASFALISMENOS] ([AMA])
GO
ALTER TABLE [dbo].[PINAKAS_METAVOLWN]  WITH CHECK ADD  CONSTRAINT
[FK_PINAKAS_METAVOLWN_OIKONOMIKH_DRASTHRIOTITA] FOREIGN
KEY([OIKONOMIKH_DRASTHRIOTHTA_ID])
REFERENCES [dbo].[OIKONOMIKH_DRASTHRIOTITA]
([OIKONOMIKH_DRASTHRIOTHTA_ID])

CREATE TABLE [dbo].[ASFALISTIKOS_FOREAS](
	[AMA] [nchar](9) COLLATE Greek_CI_AS NOT NULL,
	[ASF_FOREAS_ID] [nchar](10) COLLATE Greek_CI_AS NOT NULL,
	[FLAG_EN] [nchar](1) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_ASFALISTIKOS_FOREAS] PRIMARY KEY CLUSTERED
(
	[AMA] ASC,
	[ASF_FOREAS_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

GO
USE [IKA]
GO
ALTER TABLE [dbo].[ASFALISTIKOS_FOREAS]  WITH CHECK ADD  CONSTRAINT
[FK_ASFALISTIKOS_FOREAS_ALLOS_ASF_FOREAS] FOREIGN KEY([ASF_FOREAS_ID])
REFERENCES [dbo].[ALLOS_ASF_FOREAS] ([ASF_FOREAS_ID])
GO
ALTER TABLE [dbo].[ASFALISTIKOS_FOREAS]  WITH CHECK ADD  CONSTRAINT
[FK_ASFALISTIKOS_FOREAS_ASFALISMENOS] FOREIGN KEY([AMA])
REFERENCES [dbo].[ASFALISMENOS] ([AMA])

CREATE TABLE [dbo].[ALLOS_ASF_FOREAS](
	[ASF_FOREAS_ID] [nchar](10) COLLATE Greek_CI_AS NOT NULL,
	[DESCRIPTION] [nchar](20) COLLATE Greek_CI_AS NULL,
	[COUNTRY] [nchar](150) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_ALLOS_ASF_FOREAS] PRIMARY KEY CLUSTERED
(
	[ASF_FOREAS_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[OIKONOMIKH_DRASTHRIOTHTA](
	[OIKONOMIKH_DRASTHRIOTHTA_ID] [int] NOT NULL,
	[DESCRIPTION] [nchar](150) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_OIKONOMIKH_DRASTHRIOTHTA] PRIMARY KEY CLUSTERED
```

```
(
	[OIKONOMIKH_DRASTHRIOTHTA_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[NOMOI](
	[NOMOS_ID] [int] NOT NULL,
	[DESCRIPTION] [nchar](50) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_NOMOI] PRIMARY KEY CLUSTERED
(
	[NOMOS_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[NATIONALITY](
	[NATIONALITY_ID] [nchar](2) COLLATE Greek_CI_AS NOT NULL,
	[DESCRIPTION] [nchar](150) COLLATE Greek_CI_AS NULL,
	[NATION_DESCRIPTION] [nchar](20) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_NATIONALITY] PRIMARY KEY CLUSTERED
(
	[NATIONALITY_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

A.2 SQL Scripts δημιουργίας των πινάκων της Αποθήκης Δεδομένων IKA_OAEE_DW

```
/*Create FACT TABLE*/

CREATE TABLE [dbo].[FACT_PERIODIKH_KINHSH](
	[KWD_METANASTH] [nchar](9) COLLATE Greek_CI_AS NOT NULL,
	[KWD_PER_KATOIKIAS] [int] NULL,
	[KWD_PER_ERGASIAS] [int] NULL,
	[KWD_YPOKATASTHMATOS] [int] NULL,
	[KWD_OIKONOM_DRAST] [int] NULL,
	[KWD_HMEROM] [int] NOT NULL,
	[HMERES_ASFALISHS] [nchar](10) COLLATE Greek_CI_AS NULL,
	[APODOXES] [money] NULL,
	[EISFORES] [money] NULL,
 CONSTRAINT [PK_FACT_PERIODIKH_KINHSH] PRIMARY KEY CLUSTERED
(
	[KWD_METANASTH] ASC,
	[KWD_HMEROM] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

GO
USE [IKA_OAEE_DW]
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_GEOGRAPHY] FOREIGN
KEY([KWD_PER_KATOIKIAS])
REFERENCES [dbo].[DIM_GEOGRAPHY] ([NOMOS_ID])
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_GEOGRAPHY1] FOREIGN
KEY([KWD_PER_ERGASIAS])
```

```
REFERENCES [dbo].[DIM_GEOGRAPHY] ([NOMOS_ID])
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_METANASTHS] FOREIGN KEY([KWD_METANASTH])
REFERENCES [dbo].[DIM_METANASTHS] ([KWD_METANASTH])
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_OIKONOM_DRAST] FOREIGN
KEY([KWD_OIKONOM_DRAST])
REFERENCES [dbo].[DIM_OIKONOM_DRAST] ([KWD_OIKONOM_DRAST])
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_TIME] FOREIGN KEY([KWD_HMEROM])
REFERENCES [dbo].[DIM_TIME] ([KWD_HMEROM])
GO
ALTER TABLE [dbo].[FACT_PERIODIKH_KINHSH]  WITH CHECK ADD  CONSTRAINT
[FK_FACT_PERIODIKH_KINHSH_DIM_YPOKATAST_FOREAS] FOREIGN
KEY([KWD_YPOKATASTHMATOS])
REFERENCES [dbo].[DIM_YPOKATAST_FOREAS] ([KWD_YPOKATASTHMATOS])

/*Create DIMENSION TABLES*/

CREATE TABLE [dbo].[DIM_METANASTHS](
      [KWD_METANASTH] [nchar](9) COLLATE Greek_CI_AS NOT NULL,
      [ETOS_GENNHSHS] [int] NULL,
      [KWD_XWRA_KATAG] [nchar](2) COLLATE Greek_CI_AS NULL,
      [FYLO] [nchar](1) COLLATE Greek_CI_AS NULL,
      [ETOS_APOGRAFHS] [int] NULL,
      [KWD_OIKOG_KATAST] [nchar](10) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_METANASTHS] PRIMARY KEY CLUSTERED
(
      [KWD_METANASTH] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

GO
USE [IKA_OAEE_DW]
GO
ALTER TABLE [dbo].[DIM_METANASTHS]  WITH CHECK ADD  CONSTRAINT
[FK_DIM_METANASTHS_DIM_OIKOGEN_KATAST] FOREIGN KEY([KWD_OIKOG_KATAST])
REFERENCES [dbo].[DIM_OIKOGEN_KATAST] ([KWD_OIKOG_KATAST])
GO
ALTER TABLE [dbo].[DIM_METANASTHS]  WITH CHECK ADD  CONSTRAINT
[FK_DIM_METANASTHS_DIM_XWRA_KAT] FOREIGN KEY([KWD_XWRA_KATAG])
REFERENCES [dbo].[DIM_XWRA_KAT] ([KWD_XWRA_KATAG])

CREATE TABLE [dbo].[DIM_XWRA_KAT](
      [KWD_XWRA_KATAG] [nchar](2) COLLATE Greek_CI_AS NOT NULL,
      [DESC_XWRA_KATAG] [nchar](70) COLLATE Greek_CI_AS NULL,
      [DESC_XWRA_KATAG_ENG] [nchar](52) COLLATE Greek_CI_AS NULL,
      [ORGANISMOS] [nchar](10) COLLATE Greek_CI_AS NULL,
      [HPEIROS] [nchar](2) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_XWRA_KAT] PRIMARY KEY CLUSTERED
(
      [KWD_XWRA_KATAG] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

```
CREATE TABLE [dbo].[DIM_OIKOGEN_KATAST](
	[KWD_OIKOG_KATAST] [nchar](10) COLLATE Greek_CI_AS NOT NULL,
	[DESC_OIKOG_KATAST] [nchar](15) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_OIKOGEN_KATAST] PRIMARY KEY CLUSTERED
(
	[KWD_OIKOG_KATAST] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[DIM_TIME](
	[KWD_HMEROM] [int] IDENTITY(1,1) NOT NULL,
	[EXAMHNO] [int] NULL,
	[ETOS] [int] NULL,
 CONSTRAINT [PK_DIM_TIME] PRIMARY KEY CLUSTERED
(
	[KWD_HMEROM] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[DIM_GEOGRAPHY](
	[NOMOS_ID] [int] NOT NULL,
	[NOMOS_DESC] [nchar](20) COLLATE Greek_CI_AS NULL,
	[DIAMERISMA] [nchar](15) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_GEOGRAPHY] PRIMARY KEY CLUSTERED
(
	[NOMOS_ID] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[DIM_OIKONOM_DRAST](
	[KWD_OIKONOM_DRAST] [int] NOT NULL,
	[DESC_OIKONOM_DRAST] [nchar](150) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_OIKONOM_DRAST] PRIMARY KEY CLUSTERED
(
	[KWD_OIKONOM_DRAST] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[DIM_YPOKATAST_FOREAS](
	[KWD_YPOKATASTHMATOS] [int] NOT NULL,
	[DESC_YPOKATASTHMATOS] [nchar](60) COLLATE Greek_CI_AS NULL,
	[ASFAL_FOREAS] [nchar](10) COLLATE Greek_CI_AS NULL,
 CONSTRAINT [PK_DIM_YPOKATAST_FOREAS] PRIMARY KEY CLUSTERED
(
	[KWD_YPOKATASTHMATOS] ASC
)WITH (IGNORE_DUP_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

A.3 SQL Scripts εισαγωγής-φόρτωσης δεδομένων στους πίνακες της Αποθήκης Δεδομένων IKA_OAEE_DW

```
/*Insert data into FACT_PERIODIKH_KINHSH*/
INSERT INTO [IKA_OAEE_DW].[dbo].[FACT_PERIODIKH_KINHSH]
           ([KWD_METANASTH]
           ,[KWD_PER_KATOIKIAS]
           ,[KWD_PER_ERGASIAS]
           ,[KWD_YPOKATASTHMATOS]
           ,[KWD_OIKONOM_DRAST]
           ,[KWD_HMEROM]
           ,[HMERES_ASFALISHS]
           ,[APODOXES]
           ,[EISFORES])
SELECT a.AMA,NOMOS_ID,NOMOS_ID,'0' as KATASTHMA,
OIKONOMIKH_DRASTHRIOTHTA_ID,c.KWD_HMEROM,null,null,null
FROM IKA.dbo.ASFALISMENOS a, IKA.dbo.PINAKAS_METAVOLWN b,
IKA_OAEE_DW.dbo.DIM_TIME c
WHERE a.AMA=b.AMA
and b.ETOS=c.ETOS
and c.EXAMHNO = CASE WHEN EXAMHNO='June' THEN '1'
                WHEN EXAMHNO='December' THEN '2'
                END
ORDER BY AMA,KWD_HMEROM

/*Insert data into DIM_METANASTHS*/
INSERT INTO [IKA_OAEE_DW].[dbo].[DIM_METANASTHS]
           ([KWD_METANASTH]
           ,[ETOS_GENNHSHS]
           ,[KWD_XWRA_KATAG]
           ,[FYLO]
           ,[ETOS_APOGRAFHS]
           ,[KWD_OIKOG_KATAST])
SELECT AMA,substring(convert(varchar(10),DATE_BIRTH,101),7,4),
NATIONALITY_ID,
CASE SEX
     WHEN 'Α' THEN 'Α'
     ELSE 'Γ'
END,
ETOS_APOGRAFHS, OIKOGENEIAKH_KATAST
FROM IKA.dbo.ASFALISMENOS
UNION ALL
  SELECT AMA,ETOS_GENNISIS,XWRA_ID,FILO_ASFALISMENOU,
  substring(convert(varchar(10),HMEROMHNIA_ENARKSHS_ASFALISHS,101),7,4
  ),OIKOGENEIAKH_KATASTASH
FROM TEBE.dbo.ASFALISMENOS


/*Insert data into DIM_XWRA_KAT*/
INSERT INTO [IKA_OAEE_DW].[dbo].[DIM_XWRA_KAT]
             ([KWD_XWRA_KATAG]
             ,[DESC_XWRA_KATAG]
             ,[DESC_XWRA_KATAG_ENG]
             ,[ORGANISMOS]
             ,[HPEIROS])
```

```
SELECT
      CASE WHEN A.XWRA_ID IS NOT NULL THEN A.XWRA_ID
            ELSE NATIONALITY_ID
      END,
      CASE WHEN A.XWRA_ID IS NOT NULL THEN XWRA_EL
            ELSE RTRIM(DESCRIPTION)
      END,
      CASE WHEN A.XWRA_ID IS NOT NULL THEN XWRA_ENG
            ELSE RTRIM(DESCRIPTION)
      END,
      B.ORGANISMOS,
      A.HPEIROS
FROM TEBE.DBO.XWRA A
FULL OUTER JOIN TEBE.DBO.XWRA_ORGANISMOS b
ON (A.XWRA_ID=B.XWRA_ID)
FULL OUTER JOIN IKA.DBO.NATIONALITY C
ON (A.XWRA_ID=C.NATIONALITY_ID)

/*Insert data into DIM_GEOGRAPHY*/
INSERT INTO IKA_OAEE_DW.DBO.DIM_GEOGRAPHY
SELECT RTRIM(NOMOS_ID),RTRIM(DESCRIPTION),RTRIM(DIAMERISMA)
FROM
(SELECT A.*,B.DIAMERISMA FROM IKA.DBO.NOMOI A,TEBE.DBO.NOMOI B
WHERE A.DESCRIPTION=B.NOMOS_DESCR
UNION ALL
SELECT A.*,'ΣΤΕΡΕΑ ΕΛΛΑΔΑ' FROM IKA.DBO.NOMOI A
WHERE A.DESCRIPTION NOT IN (SELECT NOMOS_DESCR FROM TEBE.DBO.NOMOI B)
UNION ALL
SELECT 310,NOMOS_DESCR,DIAMERISMA FROM TEBE.DBO.NOMOI WHERE
NOMOS_DESCR= 'ΑΤΤΙΚΗΣ') OLIKO
ORDER BY NOMOS_ID

/*Insert data into DIM_OIKONOM_DRAST*/
INSERT INTO IKA_OAEE_DW.DBO.DIM_OIKONOM_DRAST
SELECT * FROM IKA.DBO.OIKONOMIKH_DRASTHRIOTHTA
```

A.4 SQL Scripts δημιουργίας views

```
/*Create views for OLAP analysis*/

CREATE VIEW [dbo].[PERIODIKH_KINHSH_METANASTH_OLAP]
AS
SELECT dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_KATOIKIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_ERGASIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_YPOKATASTHMATOS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM,
       dbo.FACT_PERIODIKH_KINHSH.HMERES_ASFALISHS,
       dbo.FACT_PERIODIKH_KINHSH.APODOXES,
       dbo.FACT_PERIODIKH_KINHSH.EISFORES,
       dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_GENNHSHS AS HLIKIA,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_APOGRAFHS AS
           ETH_ASFALISHS
FROM   dbo.FACT_PERIODIKH_KINHSH INNER JOIN
       dbo.DIM_METANASTHS ON
       dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH =
```

```
       dbo.DIM_METANASTHS.KWD_METANASTH
       INNER JOIN dbo.DIM_TIME ON
       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM = dbo.DIM_TIME.KWD_HMEROM


CREATE VIEW [dbo].[METANASTHS_OLAP]
AS
SELECT KWD_METANASTH, ETOS_GENNHSHS, KWD_XWRA_KATAG, FYLO,
       ETOS_APOGRAFHS, KWD_OIKOG_KATAST,
       CASE CHANGE WHEN 0 THEN 'STABLE OAEE'
       WHEN 1 THEN 'STABLE IKA' ELSE 'IKA CHANGED' END
       AS METAVOLES_OIK_DRAST
FROM
      (SELECT dbo.DIM_METANASTHS.KWD_METANASTH,
       dbo.DIM_METANASTHS.ETOS_GENNHSHS,
       dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
       dbo.DIM_METANASTHS.FYLO, dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
       dbo.DIM_METANASTHS.KWD_OIKOG_KATAST,
       COUNT(DISTINCT dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST)
       AS CHANGE
  FROM dbo.DIM_METANASTHS INNER JOIN
       dbo.FACT_PERIODIKH_KINHSH ON
       dbo.DIM_METANASTHS.KWD_METANASTH =
       dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH
  GROUP BY dbo.DIM_METANASTHS.KWD_METANASTH,
                dbo.DIM_METANASTHS.ETOS_GENNHSHS,
                dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
                dbo.DIM_METANASTHS.FYLO,
                dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
                dbo.DIM_METANASTHS.KWD_OIKOG_KATAST) AS A


/*Create views and named queries for Data mining models*/

CREATE VIEW [dbo].[PERIODIKH_KINHSH_METANASTH_DM]
AS
SELECT dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_KATOIKIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_PER_ERGASIAS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_YPOKATASTHMATOS,
       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM,
       dbo.FACT_PERIODIKH_KINHSH.HMERES_ASFALISHS,
       dbo.FACT_PERIODIKH_KINHSH.APODOXES,
       dbo.FACT_PERIODIKH_KINHSH.EISFORES,
  ISNULL(dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST, 0) AS
       KWD_OIKONOM_DRAST,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_GENNHSHS AS HLIKIA,
       dbo.DIM_TIME.ETOS - dbo.DIM_METANASTHS.ETOS_APOGRAFHS AS
           ETH_ASFALISHS
FROM   dbo.FACT_PERIODIKH_KINHSH INNER JOIN
       dbo.DIM_METANASTHS ON dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH =
       dbo.DIM_METANASTHS.KWD_METANASTH
       INNER JOIN dbo.DIM_TIME ON
       dbo.FACT_PERIODIKH_KINHSH.KWD_HMEROM = dbo.DIM_TIME.KWD_HMEROM

CREATE VIEW [dbo].[METANASTHS_DM]
AS
```

```
SELECT A.KWD_METANASTH, A.ETOS_GENNHSHS, B.DESC_XWRA_KATAG,
       B.DESC_XWRA_KATAG_ENG, B.HPEIROS, A.FYLO, A.ETOS_APOGRAFHS,
   ISNULL(C.DESC_OIKOG_KATAST, 'ΑΓΝΩΣΤΟ') AS OIKOG_KATAST,
   CASE CHANGE WHEN 0 THEN 'STABLE OAEE' WHEN 1 THEN 'STABLE IKA'
   ELSE 'IKA CHANGED' END
   AS METAVOLES_OIK_DRAST
FROM
    (SELECT dbo.DIM_METANASTHS.KWD_METANASTH,
            dbo.DIM_METANASTHS.ETOS_GENNHSHS,
            dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
            dbo.DIM_METANASTHS.FYLO,
            dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
        ISNULL(dbo.DIM_METANASTHS.KWD_OIKOG_KATAST, 0)
            AS KWD_OIKOG_KATAST,
        COUNT(DISTINCT dbo.FACT_PERIODIKH_KINHSH.KWD_OIKONOM_DRAST)
              AS CHANGE
        FROM  dbo.DIM_METANASTHS INNER JOIN
              dbo.FACT_PERIODIKH_KINHSH ON
              dbo.DIM_METANASTHS.KWD_METANASTH =
              dbo.FACT_PERIODIKH_KINHSH.KWD_METANASTH
        GROUP BY dbo.DIM_METANASTHS.KWD_METANASTH,
                 dbo.DIM_METANASTHS.ETOS_GENNHSHS,
                 dbo.DIM_METANASTHS.KWD_XWRA_KATAG,
                 dbo.DIM_METANASTHS.FYLO,
                 dbo.DIM_METANASTHS.ETOS_APOGRAFHS,
                 dbo.DIM_METANASTHS.KWD_OIKOG_KATAST)
      AS A INNER JOIN
      dbo.DIM_XWRA_KAT AS B ON A.KWD_XWRA_KATAG = B.KWD_XWRA_KATAG
      FULL OUTER JOIN
      dbo.DIM_OIKOGEN_KATAST AS C ON A.KWD_OIKOG_KATAST =
          C.KWD_OIKOG_KATAST

CREATE VIEW [dbo].[OIKONOM_DRAST_DM]
AS
SELECT [KWD_OIKONOM_DRAST]
      ,[DESC_OIKONOM_DRAST]
  FROM [IKA_OAEE_DW].[dbo].[DIM_OIKONOM_DRAST]
UNION ALL
SELECT 0 AS Expr1,'ΑΓΝΩΣΤΗ' AS Expr2

CREATE VIEW [dbo].[PLITHOS_ANA_OIKOG_KATAST] AS
SELECT * FROM (
SELECT A.OIKOG_KATAST,(PLITHOS*100/OL_PLITHOS)* PLITHOS_AGN/100 AS OLA
FROM (SELECT OIKOG_KATAST,COUNT(*) AS PLITHOS FROM METANASTHS_DM
WHERE OIKOG_KATAST <>'ΑΓΝΩΣΤΟ' GROUP BY OIKOG_KATAST
) AS A ,
(SELECT COUNT(*) AS OL_PLITHOS FROM METANASTHS_DM WHERE OIKOG_KATAST
<>'ΑΓΝΩΣΤΟ') AS B ,
(SELECT COUNT(*) AS PLITHOS_AGN FROM METANASTHS_DM WHERE
OIKOG_KATAST='ΑΓΝΩΣΤΟ') AS C ) AS D
```

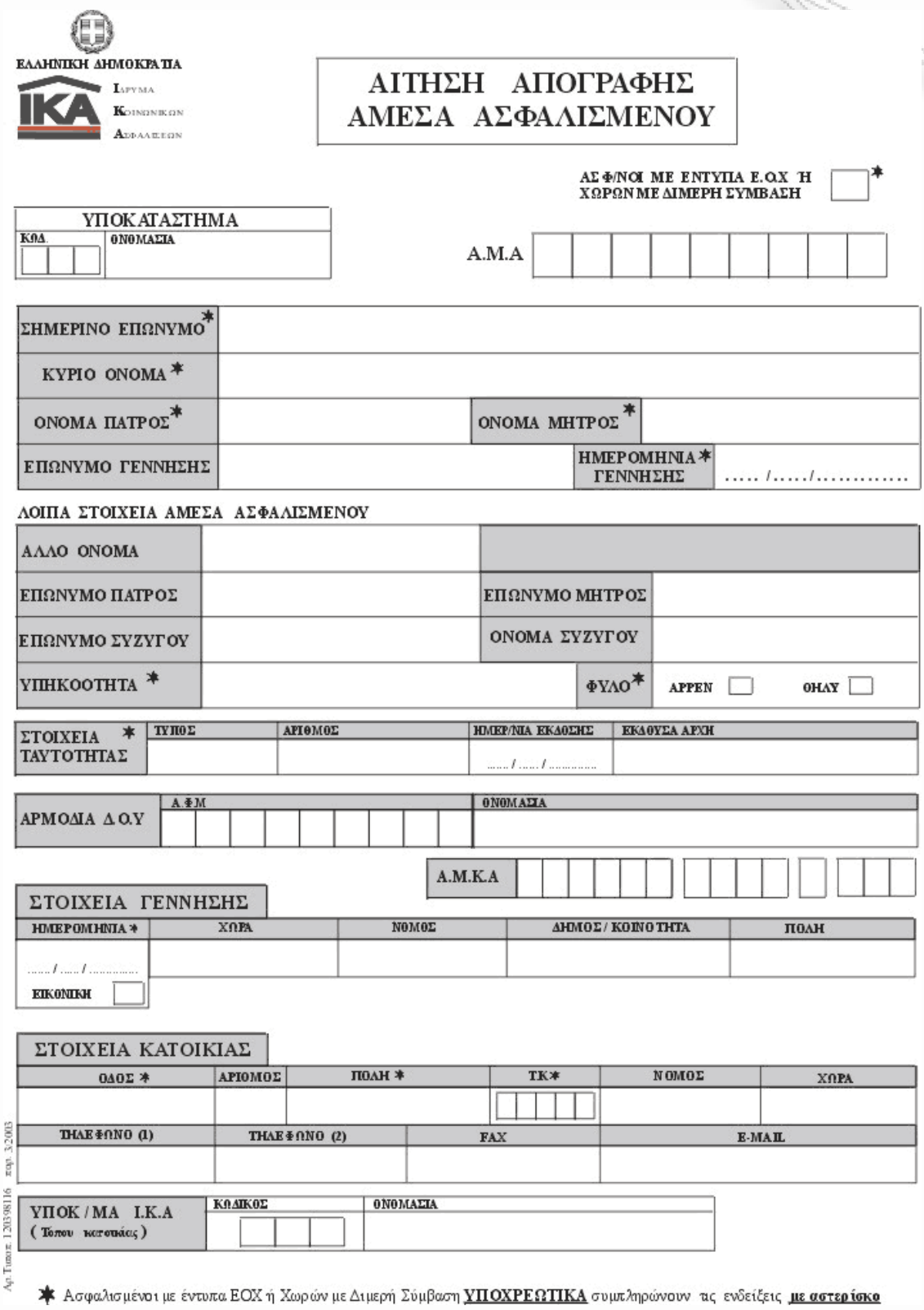
```
CREATE NAMED QUERY [dbo].[METANASTHS_DM_ADVANCED]
AS
SELECT [KWD_METANASTH]
      ,[ETOS_GENNHSHS]
      ,[DESC_XWRA_KATAG]
      ,[DESC_XWRA_KATAG_ENG]
      ,[HPEIROS]
      ,[FYLO]
      ,[ETOS_APOGRAFHS]
      ,[OIKOG_KATAST]
      ,[METAVOLES_OIK_DRAST],
CASE WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΩΣΤΟ'
THEN 0 ELSE 1 END,KWD_METANASTH)<34706 THEN 'ΕΓΓΑΜΟΣ'
WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΩΣΤΟ'
THEN 0 ELSE 1 END,KWD_METANASTH) BETWEEN 34706 AND 62864 THEN 'ΑΓΑΜΟΣ'
WHEN ROW_NUMBER() OVER (ORDER BY CASE WHEN OIKOG_KATAST='ΑΓΝΩΣΤΟ' THEN
0 ELSE 1 END,KWD_METANASTH) BETWEEN 62864 AND 64173 THEN
'ΔΙΑΖΕΥΓΜΕΝΟΣ'
ELSE OIKOG_KATAST END AS OIKOG_KATAST_2
FROM [IKA_TEBE].[dbo].[METANASTHS_DM]
```

A.5 MDX Scripts δημιουργίας μέτρων (Calculated Members) στον κύβο IKA_OAEE_DW.cube

```
[ASFALISMENOI OAEE]
sum([DIM YPOKATAST FOREAS].[ASFAL FOREAS].&[OAEE],
[Measures].[UNIQUE METANASTHS])

[ASFALISMENOI IKA]
sum([DIM YPOKATAST FOREAS].[ASFAL FOREAS].&[IKA ETAM],
[Measures].[UNIQUE METANASTHS])

[MO HLIKIAS] // Υπολογισμός μέσου όρου ηλικίας όλων των ασφαλισμένων
[Measures].[SUM HLIKIA]/[Measures].[PERIODIKH KINHSH METANASTH OLAP
Count]

[MO EISFORWN]// Υπολογισμός μέσου όρου εισφορών ανά έτος για τους
ασφαλισμένους του ΟΑΕΕ
SUM([DIM TIME].[ETOS].CURRENTMEMBER,[Measures].[EISFORES])/SUM([DIM
TIME].[ETOS].CURRENTMEMBER,[Measures].[ASFALISMENOI OAEE])

[POSOSTO METANASTWN ANA OIKON DRAST]
([Measures].[ASFALISMENOI IKA],[DIM OIKONOM DRAST].[DESC OIKONOM
DRAST].CURRENTMEMBER)/([Measures].[ASFALISMENOI IKA],[DIM OIKONOM
DRAST].[DESC OIKONOM DRAST].[ALL])

[POSOSTO METANASTWN ANA ETHNIKOTHTA]
([Measures].[METANASTHS OLAP Count],[DIM XWRA KAT].[DESC XWRA
KATAG].CurrentMember)/([Measures].[METANASTHS OLAP Count],[DIM XWRA
KAT].[DESC XWRA KATAG].[ALL])
```

Β. Δημοσιευμένα Έντυπα

Β.1 Έντυπο αίτησης απογραφής άμεσα ασφαλισμένου στο ΙΚΑ_ETAM



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΙΚΑ Ίδρυμα Κοινωνικών Ασφαλίσεων

ΑΙΤΗΣΗ ΑΠΟΓΡΑΦΗΣ
ΑΜΕΣΑ ΑΣΦΑΛΙΣΜΕΝΟΥ

ΑΣΦ/ΝΟΙ ΜΕ ΕΝΤΥΠΑ Ε.Ο.Χ Ή ΧΩΡΩΝ ΜΕ ΔΙΜΕΡΗ ΣΥΜΒΑΣΗ *

ΥΠΟΚΑΤΑΣΤΗΜΑ
ΚΩΔ. | ΟΝΟΜΑΣΙΑ

Α.Μ.Α

ΣΗΜΕΡΙΝΟ ΕΠΩΝΥΜΟ *
ΚΥΡΙΟ ΟΝΟΜΑ *
ΟΝΟΜΑ ΠΑΤΡΟΣ * | ΟΝΟΜΑ ΜΗΤΡΟΣ *
ΕΠΩΝΥΜΟ ΓΕΝΝΗΣΗΣ | ΗΜΕΡΟΜΗΝΙΑ ΓΕΝΝΗΣΗΣ * ...../...../.....

ΛΟΙΠΑ ΣΤΟΙΧΕΙΑ ΑΜΕΣΑ ΑΣΦΑΛΙΣΜΕΝΟΥ

ΑΛΛΟ ΟΝΟΜΑ
ΕΠΩΝΥΜΟ ΠΑΤΡΟΣ | ΕΠΩΝΥΜΟ ΜΗΤΡΟΣ
ΕΠΩΝΥΜΟ ΣΥΖΥΓΟΥ | ΟΝΟΜΑ ΣΥΖΥΓΟΥ
ΥΠΗΚΟΟΤΗΤΑ * | ΦΥΛΟ * ΑΡΡΕΝ ☐ ΘΗΛΥ ☐

ΣΤΟΙΧΕΙΑ ΤΑΥΤΟΤΗΤΑΣ * | ΤΥΠΟΣ | ΑΡΙΘΜΟΣ | ΗΜΕΡ/ΝΙΑ ΕΚΔΟΣΗΣ | ΕΚΔΟΥΣΑ ΑΡΧΗ

ΑΡΜΟΔΙΑ Δ.Ο.Υ | Α.Φ.Μ | ΟΝΟΜΑΣΙΑ

Α.Μ.Κ.Α

ΣΤΟΙΧΕΙΑ ΓΕΝΝΗΣΗΣ
ΗΜΕΡΟΜΗΝΙΑ * | ΧΩΡΑ | ΝΟΜΟΣ | ΔΗΜΟΣ / ΚΟΙΝΟΤΗΤΑ | ΠΟΛΗ
...../...../.....
ΕΙΚΟΝΙΚΗ ☐

ΣΤΟΙΧΕΙΑ ΚΑΤΟΙΚΙΑΣ
ΟΔΟΣ * | ΑΡΙΘΜΟΣ | ΠΟΛΗ * | ΤΚ * | ΝΟΜΟΣ | ΧΩΡΑ
ΤΗΛΕΦΩΝΟ (1) | ΤΗΛΕΦΩΝΟ (2) | FAX | E-MAIL

ΥΠΟΚ / ΜΑ Ι.Κ.Α (Τόπου κατοικίας) | ΚΩΔΙΚΟΣ | ΟΝΟΜΑΣΙΑ

Αρ. Τύπου: 1203981/6 παρ. 3/2003

* Ασφαλισμένοι με έντυπα ΕΟΧ ή Χωρών με Διμερή Σύμβαση ΥΠΟΧΡΕΩΤΙΚΑ συμπληρώνουν τις ενδείξεις με αστερίσκο

ΣΤΟΙΧΕΙΑ ΕΡΓΟΔΟΤΗ (κατά την εγγραφή)	Α.Μ.Ε ΕΡΓΟΔΟΤΗ	
	ΕΠΩΝΥΜΙΑ ΕΡΓΟΔΟΤΗ	.....
		.....
	ΤΑΧ. ΔΙΕΥΘΥΝΣΗ	.....

ΣΤΟΙΧΕΙΑ ΛΟΓΑΡΙΑΣΜΟΥ ΤΡΑΠΕΖΗΣ		ΚΩΔ ΥΠΟΚ/ΤΟΣ ΤΡΑΠΕΖΑΣ	ΟΝΟΜΑΣΙΑ ΥΠΟΚ/ΤΟΣ
ΟΝΟΜΑΣΙΑ ΤΡΑΠΕΖΑΣ	ΚΩΔ.		
		ΑΡΙΘΜΟΣ ΛΟΓΑΡΙΑΣΜΟΥ	
		.....	

ΗΜΕΡΟΜΗΝΙΑ ΕΝΑΡΞΗΣ ΑΣΦΑΛΙΣΗΣ ΣΤΟ ΙΚΑ	...../...../.....

ΑΣΦΑΛΙΣΗ ΣΕ ΑΛΛΟΥΣ ΦΟΡΕΙΣ (ΕΛΛΑΔΑ/ΑΛΛΟΔΑΠΗ) ΑΡΙΘ. ΕΙΤΡΑΦΩΝ

Α/Α	ΧΩΡΑ	ΚΩΔ.	ΦΟΡΕΑΣ	ΗΜΕΡ/ΝΙΑ ΕΝΑΡΞΗΣ	ΑΡΙΘ. ΕΘΝΙΚΟΥ ΜΗΤΡΩΟΥ	Α.Μ ΑΣΦΑΛΙΣΜΕΝΟΥ ΣΤΟΝ ΦΟΡΕΑ
				...../...../.....		
				...../...../.....		
				...../...../.....		
				...../...../.....		
				...../...../.....		
				...../...../.....		

Δηλώνεται υπεύθυνα
η ορθότητα των ανωτέρω στοιχείων

Ο / Η ΑΙΤ.....

(Υπογραφή)

ΗΜΕΡΟΜΗΝΙΑ ...../...../ 200...

ΤΟ ΜΗΤΡΩΟ ΑΣΦΑΛΙΣΜΕΝΩΝ

(Σφραγίδα - Υπογραφή)

Β.2 Στατιστική Ταξινόμηση των Κλάδων Οικονομικής Δραστηριότητας – ΣΤΑΚΟΔ 2003

(αρχείο kps_2785_6.pdf στο συνοδευτικό CD)

ΑΝΑΛΥΤΙΚΗ ΚΑΤΑΤΑΞΗ ΤΩΝ ΚΛΑΔΩΝ ΟΙΚΟΝΟΜΙΚΗΣ ΔΡΑΣΤΗΡΙΟΤΗΤΑΣ

Κωδικός	
	Α. ΓΕΩΡΓΙΑ, ΚΤΗΝΟΤΡΟΦΙΑ, ΘΗΡΑ ΚΑΙ ΔΑΣΟΚΟΜΙΑ
01	Γεωργία, κτηνοτροφία, θήρα και συναφείς βοηθητικές δραστηριότητες
02	Δασοκομία, υλοτομία και συναφείς δραστηριότητες
	Β. ΑΛΙΕΙΑ
05	Αλιεία, ιχθυοκαλλιέργεια και συναφείς βοηθητικές δραστηριότητες
	Γ. ΟΡΥΧΕΙΑ ΚΑΙ ΛΑΤΟΜΕΙΑ
	ΓΑ. ΕΞΟΡΥΞΗ ΚΑΙ ΛΑΤΟΜΗΣΗ ΕΝΕΡΓΕΙΑΚΩΝ ΥΛΙΚΩΝ
10	Εξόρυξη άνθρακα και λιγνίτη· εξόρυξη τύρφης
11	Άντληση αργού πετρελαίου και φυσικού αερίου· βοηθητικές δραστηριότητες συναφείς με την άντληση πετρελαίου και φυσικού αερίου, με εξαίρεση τις μελέτες
12	Εξόρυξη μεταλλευμάτων ουρανίου και θορίου
	ΓΒ. ΕΞΟΡΥΞΗ ΚΑΙ ΛΑΤΟΜΗΣΗ ΜΗ ΕΝΕΡΓΕΙΑΚΩΝ ΥΛΙΚΩΝ
13	Εξόρυξη μεταλλικών μεταλλευμάτων
14	Άλλες εξορυκτικές και λατομικές δραστηριότητες
	Δ. ΜΕΤΑΠΟΙΗΤΙΚΕΣ ΒΙΟΜΗΧΑΝΙΕΣ
	ΔΑ. ΒΙΟΜΗΧΑΝΙΑ ΤΡΟΦΙΜΩΝ, ΠΟΤΩΝ ΚΑΙ ΚΑΠΝΟΒΙΟΜΗΧΑΝΙΑ
15	Βιομηχανία τροφίμων και ποτών
16	Παραγωγή προϊόντων καπνού
	ΔΒ. ΠΑΡΑΓΩΓΗ ΚΛΩΣΤΟΫΦΑΝΤΟΥΡΓΙΚΩΝ ΥΛΩΝ ΚΑΙ ΠΡΟΪΟΝΤΩΝ
17	Παραγωγή κλωστοϋφαντουργικών υλών
18	Κατασκευή ειδών ένδυσης· κατεργασία και βαφή γουναρικών
	ΔΓ. ΒΙΟΜΗΧΑΝΙΑ ΔΕΡΜΑΤΟΣ ΚΑΙ ΔΕΡΜΑΤΙΝΩΝ ΕΙΔΩΝ
19	Κατεργασία και δέψη δέρματος· κατασκευή ειδών ταξιδίου (αποσκευών), τσαντών, ειδών σελοποιίας, ειδών σαγματοποιίας και υποδημάτων
	ΔΔ. ΒΙΟΜΗΧΑΝΙΑ ΞΥΛΟΥ ΚΑΙ ΠΡΟΪΟΝΤΩΝ ΞΥΛΟΥ
20	Βιομηχανία ξύλου και κατασκευή προϊόντων από ξύλο και φελλό, εκτός από τα έπιπλα· κατασκευή ειδών καλαθοποιίας και σπαρτοπλεκτικής
	ΔΕ. ΠΑΡΑΓΩΓΗ ΧΑΡΤΟΠΟΛΤΟΥ, ΠΑΡΑΓΩΓΗ ΧΑΡΤΙΟΥ ΚΑΙ ΠΡΟΪΟΝΤΩΝ ΑΠΟ ΧΑΡΤΙ· ΕΚΔΟΤΙΚΕΣ ΚΑΙ ΕΚΤΥΠΩΤΙΚΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ
21	Παραγωγή χαρτοπολτού, χαρτιού και προϊόντων από χαρτί
22	Εκδόσεις, εκτυπώσεις και αναπαραγωγή προεγγεγραμμένων μέσων εγγραφής ήχου και εικόνας και μέσων πληροφορικής
	ΔΣΤ. ΠΑΡΑΓΩΓΗ ΟΠΤΑΝΘΡΑΚΑ (ΚΩΚ), ΠΡΟΪΟΝΤΩΝ ΔΙΥΛΙΣΗΣ ΠΕΤΡΕΛΑΙΟΥ ΚΑΙ ΠΥΡΗΝΙΚΩΝ ΚΑΥΣΙΜΩΝ
23	Παραγωγή οπτάνθρακα (κωκ), προϊόντων διύλισης πετρελαίου και πυρηνικών καυσίμων
	ΔΖ. ΠΑΡΑΓΩΓΗ ΧΗΜΙΚΩΝ ΟΥΣΙΩΝ, ΧΗΜΙΚΩΝ ΠΡΟΪΟΝΤΩΝ ΚΑΙ ΣΥΝΘΕΤΙΚΩΝ ΙΝΩΝ
24	Παραγωγή χημικών ουσιών και προϊόντων
	ΔΗ. ΚΑΤΑΣΚΕΥΗ ΠΡΟΪΟΝΤΩΝ ΑΠΟ ΕΛΑΣΤΙΚΟ (ΚΑΟΥΤΣΟΥΚ) ΚΑΙ ΠΛΑΣΤΙΚΕΣ ΥΛΕΣ
25	Κατασκευή προϊόντων από ελαστικό (καουτσούκ) και πλαστικές ύλες
	ΔΘ. ΚΑΤΑΣΚΕΥΗ ΑΛΛΩΝ ΠΡΟΪΟΝΤΩΝ ΑΠΟ ΜΗ ΜΕΤΑΛΛΙΚΑ ΟΡΥΚΤΑ
26	Κατασκευή άλλων προϊόντων από μη μεταλλικά ορυκτά
	ΔΙ. ΠΑΡΑΓΩΓΗ ΒΑΣΙΚΩΝ ΜΕΤΑΛΛΩΝ ΚΑΙ ΚΑΤΑΣΚΕΥΗ ΜΕΤΑΛΛΙΚΩΝ ΠΡΟΪΟΝΤΩΝ
27	Παραγωγή βασικών μετάλλων
28	Κατασκευή μεταλλικών προϊόντων, με εξαίρεση τα μηχανήματα και τα είδη εξοπλισμού
	ΔΚ. ΚΑΤΑΣΚΕΥΗ ΜΗΧΑΝΗΜΑΤΩΝ ΚΑΙ ΕΙΔΩΝ ΕΞΟΠΛΙΣΜΟΥ Μ.Α.Κ.
29	Κατασκευή μηχανημάτων και ειδών εξοπλισμού μ.α.κ.
	ΔΛ. ΚΑΤΑΣΚΕΥΗ ΗΛΕΚΤΡΙΚΟΥ ΕΞΟΠΛΙΣΜΟΥ ΚΑΙ ΟΠΤΙΚΩΝ ΣΥΣΚΕΥΩΝ
30	Κατασκευή μηχανών γραφείου και ηλεκτρονικών υπολογιστών
31	Κατασκευή ηλεκτρικών μηχανών και συσκευών μ.α.κ.

Κωδικός		
32		Κατασκευή εξοπλισμού και συσκευών ραδιοφωνίας, τηλεόρασης και επικοινωνιών
33		Κατασκευή ιατρικών οργάνων, οργάνων ακριβείας και οπτικών οργάνων, κατασκευή φολογιών κάθε είδους
	ΔΜ.	ΚΑΤΑΣΚΕΥΗ ΕΞΟΠΛΙΣΜΟΥ ΜΕΤΑΦΟΡΩΝ
34		Κατασκευή αυτοκίνητων οχημάτων· κατασκευή ρυμουλκούμενων και ημιρυμουλκούμενων οχημάτων
35		Κατασκευή λοιπού εξοπλισμού μεταφορών
	ΔΝ.	ΛΟΙΠΕΣ ΒΙΟΜΗΧΑΝΙΕΣ Μ.Α.Κ.
36		Κατασκευή επίπλων· λοιπές βιομηχανίες μ.α.κ.
37		Ανακύκλωση
	Ε.	**ΠΑΡΟΧΗ ΗΛΕΚΤΡΙΚΟΥ ΡΕΥΜΑΤΟΣ, ΦΥΣΙΚΟΥ ΑΕΡΙΟΥ ΚΑΙ ΝΕΡΟΥ**
40		Παροχή ηλεκτρικού ρεύματος, φυσικού αερίου, ατμού και ζεστού νερού
41		Συλλογή, καθαρισμός και διανομή νερού
	ΣΤ.	**ΚΑΤΑΣΚΕΥΕΣ**
45		Κατασκευές
	Ζ.	**ΧΟΝΔΡΙΚΟ ΚΑΙ ΛΙΑΝΙΚΟ ΕΜΠΟΡΙΟ· ΕΠΙΣΚΕΥΗ ΑΥΤΟΚΙΝΗΤΩΝ ΟΧΗΜΑΤΩΝ, ΜΟΤΟΣΥΚΛΕΤΩΝ ΚΑΙ ΕΙΔΩΝ ΠΡΟΣΩΠΙΚΗΣ ΚΑΙ ΟΙΚΙΑΚΗΣ ΧΡΗΣΗΣ**
50		Εμπόριο, συντήρηση και επισκευή αυτοκινήτων οχημάτων και μοτοσυκλετών· λιανική πώληση καυσίμων οχημάτων
51		Χονδρικό εμπόριο και εμπόριο με προμήθεια, εκτός από το εμπόριο αυτοκίνητων οχημάτων και μοτοσυκλετών
52		Λιανικό εμπόριο, εκτός από το εμπόριο αυτοκίνητων οχημάτων και μοτοσυκλετών· επισκευή ειδών ατομικής και οικιακής χρήσης
	Η.	**ΞΕΝΟΔΟΧΕΙΑ ΚΑΙ ΕΣΤΙΑΤΟΡΙΑ**
55		Ξενοδοχεία και εστιατόρια
	Θ.	**ΜΕΤΑΦΟΡΕΣ, ΑΠΟΘΗΚΕΥΣΗ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΕΣ**
60		Χερσαίες μεταφορές. Μεταφορές μέσω αγωγών
61		Υδάτινες μεταφορές
62		Εναέριες μεταφορές
63		Βοηθητικές και συναφείς προς τις μεταφορές δραστηριότητες· δραστηριότητες ταξιδιωτικών πρακτορείων
64		Ταχυδρομεία και τηλεπικοινωνίες
	Ι.	**ΕΝΔΙΑΜΕΣΟΙ ΧΡΗΜΑΤΟΠΙΣΤΩΤΙΚΟΙ ΟΡΓΑΝΙΣΜΟΙ**
65		Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί, με εξαίρεση τις ασφαλιστικές εταιρείες και τα ταμεία συντάξεων
66		Ασφαλίσεις και συνταξιοδοτικά ταμεία, εκτός από την υποχρεωτική κοινωνική ασφάλιση
67		Δραστηριότητες συναφείς με τις δραστηριότητες ενδιάμεσων χρηματοπιστωτικών οργανισμών
	Κ.	**ΔΙΑΧΕΙΡΙΣΗ ΑΚΙΝΗΤΗΣ ΠΕΡΙΟΥΣΙΑΣ, ΕΚΜΙΣΘΩΣΕΙΣ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ**
70		Δραστηριότητες σχετικές με ακίνητη περιουσία
71		Εκμίσθωση μηχανημάτων και εξοπλισμού χωρίς χειριστή· εκμίσθωση ειδών ατομικής και οικιακής χρήσης
72		Πληροφορική και συναφείς δραστηριότητες
73		Έρευνα και ανάπτυξη
74		Άλλες επιχειρηματικές δραστηριότητες
	Λ.	**ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ ΚΑΙ ΑΜΥΝΑ· ΥΠΟΧΡΕΩΤΙΚΗ ΚΟΙΝΩΝΙΚΗ ΑΣΦΑΛΙΣΗ**
75		Δημόσια διοίκηση και άμυνα· υποχρεωτική κοινωνική ασφάλιση

Κωδικός	
	Μ. ΕΚΠΑΙΔΕΥΣΗ
80	Εκπαίδευση
	Ν. ΥΓΕΙΑ ΚΑΙ ΚΟΙΝΩΝΙΚΗ ΜΕΡΙΜΝΑ
85	Υγεία και κοινωνική μέριμνα
	Ξ. ΑΛΛΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ ΠΑΡΟΧΗΣ ΥΠΗΡΕΣΙΩΝ ΥΠΕΡ ΤΟΥ ΚΟΙΝΩΝΙΚΟΥ ΣΥΝΟΛΟΥ ΚΑΙ ΑΛΛΩΝ ΥΠΗΡΕΣΙΩΝ ΚΟΙΝΩΝΙΚΟΥ Ή ΑΤΟΜΙΚΟΥ ΧΑΡΑΚΤΗΡΑ
90	Διάθεση λυμάτων και απορριμμάτων· υγιεινή και παρόμοιες δραστηριότητες
91	Δραστηριότητες οργανώσεων με μέλη μ.α.κ.
92	Ψυχαγωγικές, πολιτιστικές και αθλητικές δραστηριότητες
93	Άλλες δραστηριότητες παροχής υπηρεσιών
	Ο. ΙΔΙΩΤΙΚΑ ΝΟΙΚΟΚΥΡΙΑ ΠΟΥ ΑΠΑΣΧΟΛΟΥΝ ΟΙΚΙΑΚΟ ΠΡΟΣΩΠΙΚΟ ΚΑΙ ΜΗ ΔΙΑΦΟΡΟΠΟΙΗΜΕΝΕΣ ΠΑΡΑΓΩΓΙΚΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ ΝΟΙΚΟΚΥΡΙΩΝ ΓΙΑ ΙΔΙΑ ΧΡΗΣΗ
95	Δραστηριότητες νοικοκυριών ως εργοδοτών οικιακού προσωπικού
96	Μη διαφοροποιημένες δραστηριότητες ιδιωτικών νοικοκυριών, που αφορούν στην παραγωγή αγαθών για ίδια χρήση
97	Μη διαφοροποιημένες δραστηριότητες ιδιωτικών νοικοκυριών, που αφορούν στην παραγωγή υπηρεσιών για ίδια χρήση
	Π. ΕΤΕΡΟΔΙΚΟΙ ΟΡΓΑΝΙΣΜΟΙ ΚΑΙ ΟΡΓΑΝΑ
99	Ετερόδικοι οργανισμοί και όργανα

Β.3 Εξαμηνιαία Στατιστικά Στοιχεία Απασχόλησης χρονικής περιόδου Ιούνιος 2004 έως Ιούνιος 2008

Στη συνέχεια ακολουθούν εννέα πίνακες με τα εξαμηνιαία στατιστικά στοιχεία απασχόλησης των ασφαλισμένων του ΙΚΑ για τη χρονική περίοδο Ιούνιος 2004 έως Ιούνιος 2008 τα οποία υπάρχουν στο site www.ika.gr και τα αντίστοιχα αρχεία δίνονται στο φάκελο IKA_Statistika στο συνοδευτικό CD.

ΙΟΥΝΙΟΣ 2004

ΠΙΝΑΚΑΣ 6
ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ

Οικονομική δραστηριότητα		Έλληνες	%	Χώρες Ε.Ε. (24)	%	Αλβανοί	%	Υπόλοιποι	%	Σύνολο	%
Γεωργία, Κτηνοτροφία Θήρα και Δασοκομία	Α	6346	0,39	24	0,18	703	0,53	555	0,57	7628	0,40
Αλιεία	Β	2564	0,16	18	0,14	179	0,13	329	0,34	3090	0,16
Ορυχεία και Λατομεία	Γ	7507	0,46	24	0,18	308	0,23	254	0,26	8093	0,43
Μεταποιητικές Βιομηχανίες	Δ	331385	20,15	1246	9,43	21378	16,12	23612	24,17	377621	20,00
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	15207	0,92	7	0,05	107	0,08	65	0,07	15386	0,81
Κατασκευές	ΣΤ	177980	10,82	1615	12,23	61591	46,43	21055	21,55	262241	13,89
Χονδρικό και Λιανικό εμπόριο	Ζ	355290	21,60	1756	13,29	15354	11,57	12010	12,29	384410	20,36
Ξενοδοχεία και Εστιατόρια	Η	155809	9,47	3422	25,90	19405	14,63	16150	16,53	194786	10,32
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	110038	6,69	1118	8,46	1779	1,34	1972	2,02	114907	6,09
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	55099	3,35	283	2,14	110	0,08	189	0,19	55681	2,95
Διαχείριση ακίνητης περιουσίας	Κ	114750	6,98	1613	12,21	4306	3,25	4046	4,14	124715	6,61
Δημόσια διοίκηση και άμυνα	Λ	55721	3,39	61	0,46	73	0,06	73	0,07	55928	2,96
Εκπαίδευση	Μ	79983	4,86	638	4,83	675	0,51	681	0,70	81977	4,34
Υγεία και κοινωνική μέριμνα	Ν	66590	4,05	419	3,17	841	0,63	837	0,86	68687	3,64
Άλλες δραστηριότητες	Ξ	92029	5,60	528	4,00	2651	2,00	2212	2,26	97420	5,16
Ιδιωτικά νοικοκυριά	Ο	6064	0,37	305	2,31	2673	2,01	13071	13,38	22113	1,17
Ετερόδικοι οργανισμοί	Π	1275	0,08	45	0,34	7	0,01	46	0,05	1373	0,07
Αγνωστο		10953	0,67	88	0,67	516	0,39	539	0,55	12096	0,64
Σύνολο		**1.644.590**	**100,00**	**13.210**	**100,00**	**132.656**	**100,00**	**97.696**	**100,00**	**1888152**	**100,00**

* Ο αριθμός των ασφαλισμένων είναι διακριτός

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗ ΜΕΓΑΛΥΤΕΡΗ ΣΥΧΝΟΤΗΤΑ ΑΠΑΣΧΟΛΗΣΗΣ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρικό εμπόριο 20,36%, στις μεταποιητικές το 20%,στις κατασκευές 13,89%.
2. Οι Έλληνες απασχολούνται στο χονδρικό εμπόριο 21,60% στις μεταποιητικές 20,15% και στις κατασκευές 10,82%.
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στα ξενοδοχεία 25,90%, στο χονδρικό και λιανικό εμπόριο 13,29% & στις κατασκευές 12,23%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 46,43%, στις μεταποιητικές βιομηχανίες 16,12% και στα ξενοδοχεία 14,63%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές 24,17%, στις κατασκευές 21,55% και στα ξενοδοχεία 16,53% .

ΔΕΚΕΜΒΡΙΟΣ 2004

ΠΙΝΑΚΑΣ 6
ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ

Οικονομική δραστηριότητα		Έλληνες	%	Χώρες Ε.Ε. (24)	%	Αλβανοί	%	Υπόλοιποι	%	Σύνολο	%
Γεωργία, Κτηνοτροφία Θήρα και Δασοκομία	Α	7.409	0,47	17	0,18	537	0,48	389	0,45	8.352	0,47
Αλιεία	Β	2.313	0,15	14	0,15	127	0,11	221	0,26	2.675	0,15
Ορυχεία και Λατομεία	Γ	6.664	0,42	22	0,23	274	0,24	206	0,24	7.166	0,40
Μεταποιητικές Βιομηχανίες	Δ	318.513	20,22	1.233	12,95	20.327	18,16	22.648	26,22	362.721	20,34
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	13.140	0,83	6	0,06	85	0,08	52	0,06	13.283	0,74
Κατασκευές	ΣΤ	160.729	10,20	1.545	16,23	52.026	46,48	18.443	21,35	232.743	13,05
Χονδρικό και Λιανικό εμπόριο	Ζ	357.429	22,69	1.500	15,75	14.849	13,27	11.625	13,46	385.403	21,61
Ξενοδοχεία και Εστιατόρια	Η	93.187	5,92	953	10,01	10.098	9,02	9.960	11,53	114.198	6,40
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	118.975	7,55	530	5,57	1.784	1,59	1.832	2,12	123.121	6,90
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	57.895	3,68	292	3,07	151	0,13	205	0,24	58.543	3,28
Διαχείριση ακίνητης περιουσίας	Κ	112.482	7,14	938	9,85	4.165	3,72	3.883	4,50	121.468	6,81
Δημόσια διοίκηση και άμυνα	Λ	49.836	3,16	55	0,58	92	0,08	67	0,08	50.050	2,81
Εκπαίδευση	Μ	101.281	6,43	1.068	11,22	786	0,70	803	0,93	103.938	5,83
Υγεία και κοινωνική μέριμνα	Ν	68.271	4,33	435	4,57	871	0,78	827	0,96	70.404	3,95
Άλλες δραστηριότητες	Ξ	91.148	5,79	485	5,09	2.553	2,28	2.049	2,37	96.235	5,40
Ιδιωτικά νοικοκυριά	Ο	6.259	0,40	307	3,22	2.740	2,45	12.600	14,59	21.906	1,23
Ετερόδικοι οργανισμοί	Π	1.277	0,08	41	0,43	10	0,01	47	0,05	1.375	0,08
Αγνωστο		8.525	0,54	81	0,85	450	0,40	527	0,61	9.583	0,54
Σύνολο		**1.575.333**	**100,00**	**9.522**	**100,00**	**111.925**	**100,00**	**86.384**	**100,00**	**1.783.164**	**100,00**

* Ο αριθμός των ασφαλισμένων είναι διακριτός

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗ ΜΕΓΑΛΥΤΕΡΗ ΣΥΧΝΟΤΗΤΑ ΑΠΑΣΧΟΛΗΣΗΣ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρικό εμπόριο 21,61%, στις μεταποιητικές το 20,34%,στις κατασκευές 13,05%.
2. Οι Έλληνες απασχολούνται στο χονδρικό εμπόριο 22,69%, στις μεταποιητικές 20,22% και στις κατασκευές 10,20%.
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στις κατασκευές 16,23%, στο χονδρικό εμπόριο 15,75% & στις μεταποιητικές βιομηχανίες 12,95%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 46,48%, στις μεταποιητικές βιομηχανίες 18,16%, στο χονδρικό και λιανικό εμπόριο 13,27% .
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές 26,22%, στις κατασκευές 21,35%, στα ιδιωτικά νοικοκυριά 14,59%.

ΙΟΥΝΙΟΣ 2005 / JUNE 2005

ΠΙΝΑΚΑΣ 6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	5.319	0,33	17	0,12	537	0,42	276	0,30	6.149	0,33	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.088	0,13	15	0,11	107	0,08	116	0,13	2.326	0,13	Β	Fishing
Ορυχεία και Λατομεία	Γ	6.856	0,42	30	0,22	255	0,20	228	0,25	7.369	0,40	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	308.986	19,09	1.250	8,97	19.411	15,13	21.353	23,42	351.000	18,95	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	10.826	0,67	4	0,03	101	0,08	56	0,06	10.987	0,59	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	162.460	10,03	1.642	11,79	58.282	45,42	19.714	21,62	242.098	13,07	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	359.309	22,19	1.877	13,47	15.352	11,96	12.118	13,29	388.656	20,98	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	155.399	9,60	3.739	26,84	20.261	15,79	16.097	17,66	195.496	10,55	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	127.278	7,86	1.214	8,71	1.879	1,46	1.948	2,14	132.319	7,14	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	52.697	3,26	296	2,12	145	0,11	190	0,21	53.328	2,88	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	114.514	7,07	1.667	11,97	4.490	3,50	4.163	4,57	124.834	6,74	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	58.166	3,59	58	0,42	113	0,09	66	0,07	58.403	3,15	Λ	Public administration & defence, compulsory socoal security
Εκπαίδευση	Μ	79.188	4,89	681	4,89	702	0,55	649	0,71	81.220	4,38	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	68.918	4,26	477	3,42	933	0,73	833	0,91	71.161	3,84	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	89.195	5,51	548	3,93	2.797	2,18	2.218	2,43	94.758	5,12	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.841	0,36	260	1,87	2.391	1,86	10.539	11,56	19.031	1,03	Ο	Private households with employed persons
Ετερόδικοι οργανισμοί	Π	1.297	0,08	53	0,38	9	0,01	45	0,05	1.404	0,08	Π	Extra territorial organizations & bodies
Αγνωστο		10.609	0,66	104	0,75	549	0,43	559	0,61	11.821	0,64		Unknown
Σύνολο		**1.618.946**	**100,00**	**13.932**	**100,00**	**128.314**	**100,00**	**91.168**	**100,00**	**1.852.360**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 20,98%, στις μεταποιητικές το 18,95%,στις κατασκευές 13,07%.
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 22,19%, στις μεταποιητικές 19,09% και στις κατασκευές 10,03%.
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στα ξενοδοχεία και εστιατόρια 26,84%, στο χονδρικό εμπόριο 13,47% & στις κατασκευές 11,79%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 45,42%, στα ξενοδοχεία και εστιατόρια 15,79% και στις μεταποιητικές βιομηχανίες 15,13%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές 23,42%, στις κατασκευές 21,62%, στα ξενοδοχεία και εστιατόρια 17,66%.

COMMENTS ON EMPLOYER´S ECONOMIC ACTIVITY

1) 20,98% of Insured Population is empoloyed in ´Wholesale & retail trade´, 18,95% in ´Manufacturing´ and 13,07% in ´Conctructions´
2) 22,19% of Insured Greeks is employed in ´Wholesale & retail trade´, 19,09% in ´Manufacturing´ and 10,03% in ´Constructions´
3) 26,84% of insured EU citizens is employed in ´Hotels and restaurants´, 13,47% in ´Wholesale & retail trade´and 8,97% in ´Manufacturing´
4) 45,42% of insured Albanians is employed in ´Constructions´, 15,79% in ´Hotels and restaurants´ and 15,13% in ´Manufacturing´
5) 23,42% of the insured foreigner workers is employed in ´Manufacturing´, 21,62% in ´Constructions´ and 17,66% in ´Hotels and restaurants´

ΔΕΚΕΜΒΡΙΟΣ 2005 / DECEMBER 2005

ΠΙΝΑΚΑΣ 6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	6.566	0,41	15	0,14	502	0,42	285	0,33	7.368	0,40	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.148	0,13	15	0,14	92	0,08	117	0,13	2.372	0,13	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.360	0,45	27	0,26	297	0,25	239	0,27	7.923	0,43	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	315.156	19,46	1.259	12,10	19.904	16,71	21.803	25,03	358.122	19,50	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	14.822	0,92	9	0,09	72	0,06	51	0,06	14.954	0,81	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	161.014	9,94	1.708	16,42	57.775	48,51	20.238	23,24	240.735	13,11	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	372.682	23,01	1.664	16,00	15.405	12,93	12.204	14,01	401.955	21,89	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	96.446	5,96	1.106	10,63	10.865	9,12	10.078	11,57	118.495	6,45	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	123.637	7,63	605	5,82	1.870	1,57	1.928	2,21	128.040	6,97	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	53.632	3,31	313	3,01	153	0,13	194	0,22	54.292	2,96	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	117.681	7,27	975	9,37	4.493	3,77	4.187	4,81	127.336	6,94	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	56.293	3,48	60	0,58	91	0,08	74	0,08	56.518	3,08	Λ	Public administration & defence, compulsory socoal security
Εκπαίδευση	Μ	105.273	6,50	1.193	11,47	848	0,71	845	0,97	108.159	5,89	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	73.552	4,54	477	4,59	1.002	0,84	831	0,95	75.862	4,13	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	95.958	5,93	563	5,41	2.753	2,31	2.249	2,58	101.523	5,53	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.837	0,36	268	2,58	2.484	2,09	11.158	12,81	19.747	1,08	Ο	Private households with employed persons
Ετερόδικοι οργανισμοί	Π	1.274	0,08	57	0,55	10	0,01	62	0,07	1.403	0,08	Π	Extra territorial organizations & bodies
Αγνωστο		10.168	0,63	87	0,84	495	0,42	558	0,64	11.308	0,62		Unknown
Σύνολο		**1.619.499**	**100,00**	**10.401**	**100,00**	**119.111**	**100,00**	**87.101**	**100,00**	**1.836.112**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ
1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 21,89%, στις μεταποιητικές το 19,50%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 23,01%, στις μεταποιητικές 19,46%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στις κατασκευές 16,42% και στο χονδρικό εμπόριο 16%
4. Οι Αλβανοί απασχολούνται στις κατασκευές 48,51% και στις μεταποιητικές βιομηχανίες 16,71%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές 25,03%, στις κατασκευές 23,24%.

COMMENTS ON EMPLOYER´S ECONOMIC ACTIVITY
1) 21,89% of Insured Population is empoloyed in ´Wholesale & retail trade´ and 19,5% in ´Manufacturing´
2) 23,01% of Insured Greeks is employed in ´Wholesale & retail trade´, 19,46% in ´Manufacturing´
3) 16,42% of insured EU citizens is employed in ´Constructions´, 16% in ´Wholesale & retail trade´
4) 48,51% of insured Albanians is employed in ´Constructions´and 16,71% in ´Manufacturing´
5) 25,03% of the insured foreigner workers is employed in ´Manufacturing´ and 23,24% in ´Constructions´

ΙΟΥΝΙΟΣ 2006 / JUNE 2006

ΠΙΝΑΚΑΣ 6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	5.529	0,33	21	0,13	587	0,40	321	0,29	6.458	0,33	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.171	0,13	17	0,10	93	0,06	118	0,11	2.399	0,12	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.395	0,44	25	0,15	315	0,21	262	0,24	7.997	0,41	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	306.970	18,14	1.337	8,20	20.298	13,85	23.883	21,64	352.488	17,93	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	15.955	0,94	8	0,05	95	0,06	60	0,05	16.118	0,82	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	169.386	10,01	2.013	12,35	69.894	47,69	25.601	23,19	266.894	13,58	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	376.442	22,25	2.142	13,14	16.730	11,41	14.837	13,44	410.151	20,87	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	160.721	9,50	4.798	29,42	22.777	15,54	19.516	17,68	207.812	10,57	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	126.153	7,57	1.353	8,30	2.112	1,44	2.394	2,17	134.012	6,82	Θ	Transport storage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	55.549	3,28	316	1,94	180	0,12	207	0,19	56.252	2,86	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	121.657	7,19	1.749	10,73	5.144	3,51	5.167	4,68	133.717	6,80	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	67.033	3,96	70	0,43	132	0,09	107	0,10	67.342	3,43	Λ	Public administration & defence, compulsory socoal security
Εκπαίδευση	Μ	85.274	5,04	834	5,11	753	0,51	696	0,63	87.557	4,45	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	74.597	4,41	519	3,18	1.046	0,71	909	0,82	77.071	3,92	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	97.384	5,75	659	4,04	3.159	2,16	2.547	2,31	103.749	5,28	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.737	0,34	266	1,63	2.593	1,77	12.913	11,70	21.509	1,09	Ο	Private households with employed persons
Ετερόδικοι οργανισμοί	Π	1.218	0,07	64	0,39	9	0,01	58	0,05	1.349	0,07	Π	Extra territorial organizations & bodies
Άγνωστο		11.077	0,65	115	0,71	654	0,45	789	0,71	12.635	0,64		Unknown
Σύνολο		**1.692.248**	**100,00**	**16.306**	**100,00**	**146.571**	**100,00**	**110.385**	**100,00**	**1.965.510**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 20,87%, στις μεταποιητικές το 17,93%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 22,25%, στις μεταποιητικές 18,14%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στα ξενοδοχεία και εστιατόρια 29,42% και στο χονδρικό εμπόριο 13,14%
4. Οι Αλβανοί απασχολούνται στις κατασκευές 47,69% και στα ξενοδοχεία και εστιατόρια 15,54%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις κατασκευές 23,19% και στις μεταποιητικές βιομηχανίες 21,64%.

COMMENTS ON EMPLOYER'S ECONOMIC ACTIVITY

1) 20,87% of Insured Population is empoloyed in 'Wholesale & retail trade' and 17,93% in 'Manufacturing'
2) 22,25% of Insured Greeks is employed in 'Wholesale & retail trade', 18,14% in 'Manufacturing'
3) 29,42% of insured EU citizens is employed in 'Hotels and restaurants', 13,14% in 'Wholesale & retail trade'
4) 47,69% of insured Albanians is employed in 'Constructions'and 15,54% in 'Hotels and restaurants'
5) 23,19% of the insured foreigner workers is employed in 'Manufacturing' and 21,64% in 'Constructions'

ΔΕΚΕΜΒΡΙΟΣ 2006 / DECEMBER 2006

ΠΙΝΑΚΑΣ 6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 6: ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	6.492	0,39	20	0,16	576	0,43	313	0,30	7.401	0,39	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.269	0,14	21	0,17	93	0,07	132	0,13	2.515	0,13	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.575	0,46	29	0,24	321	0,24	290	0,28	8.215	0,43	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	311.935	18,74	1.472	12,14	21.146	15,94	25.434	24,49	359.987	18,82	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	15.480	0,93	12	0,10	112	0,08	72	0,07	15.676	0,82	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	163.730	9,84	2.118	17,47	65.292	49,22	24.417	23,51	255.557	13,36	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	387.979	23,31	1.940	16,00	16.999	12,82	15.397	14,83	422.315	22,07	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	98.514	5,92	1.378	11,37	12.201	9,20	12.107	11,66	124.200	6,49	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	123.611	7,43	726	5,99	2.101	1,58	2.394	2,31	128.832	6,73	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	56.205	3,38	311	2,57	188	0,14	223	0,21	56.927	2,98	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	123.012	7,39	1.195	9,86	5.124	3,86	5.271	5,08	134.602	7,04	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	63.505	3,82	70	0,58	114	0,09	111	0,11	63.800	3,33	Λ	Public administration & defence, compulsory socoal security
Εκπαίδευση	Μ	109.268	6,56	1.237	10,20	924	0,70	942	0,91	112.371	5,87	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	77.960	4,68	523	4,31	1.114	0,84	906	0,87	80.503	4,21	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	97.974	5,89	621	5,12	3.030	2,28	2.517	2,42	104.142	5,44	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.893	0,35	289	2,38	2.618	1,97	12.540	12,07	21.340	1,12	Ο	Private households with employed persons
Ετερόδικοι οργανισμοί	Π	1.178	0,07	72	0,59	7	0,01	67	0,06	1.324	0,07	Π	Extra territorial organizations & bodies
Αγνωστο		12.000	0,72	90	0,74	688	0,52	723	0,70	13.501	0,71		Unknown
Σύνολο		**1.664.580**	**100,00**	**12.124**	**100,00**	**132.648**	**100,00**	**103.856**	**100,00**	**1.913.208**	**100,00**		**Total**

*Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ
1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 22,07%, στις μεταποιητικές το 18,82%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 23,31%, στις μεταποιητικές 18,74%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στις κατασκευές 17,47% και στο χονδρικό εμπόριο 16%
4. Οι Αλβανοί απασχολούνται στις κατασκευές 49,22% και στις μεταποιητικές βιομηχανίες 15,94%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές βιομηχανίες 24,49% και στις κατασκευές 23,51% .

COMMENTS ON EMPLOYER'S ECONOMIC ACTIVITY
1) 22,07% of Insured Population is empolyed in "Wholesale & retail trade" and 18,82% in "Manufacturing"
2) 23,31% of Insured Greeks is employed in "Wholesale & retail trade", 18,74% in "Manufacturing"
3) 17,47% of insured EU citizens is employed in "Constructions", 16% in "Wholesale & retail trade"
4) 49,22% of insured Albanians is employed in "Constructions"and 15,94% in "Manufacturing"
5) 24,49% of the insured foreigner workers is employed in "Manufacturing" and 23,51% in "Constructions"

ΙΟΥΝΙΟΣ 2007 / JUNE 2007

ΠΙΝΑΚΑΣ 6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	5.513	0,31	83	0,16	654	0,42	252	0,30	6.502	0,32	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.353	0,13	44	0,09	87	0,06	95	0,11	2.579	0,13	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.628	0,43	112	0,22	327	0,21	196	0,23	8.263	0,40	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	317.436	18,07	6.912	13,63	22.064	14,30	20.696	24,25	367.108	17,93	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	16.319	0,93	19	0,04	94	0,06	44	0,05	16.476	0,80	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	164.944	9,39	10.141	20,00	70.366	45,62	17.651	20,69	263.102	12,85	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	397.400	22,62	6.062	11,95	18.212	11,81	12.876	15,09	434.550	21,23	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	165.635	9,43	13.629	26,87	25.291	16,40	13.546	15,88	218.101	10,65	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	132.712	7,55	2.162	4,26	2.377	1,54	2.213	2,59	139.464	6,81	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	59.596	3,39	366	0,72	200	0,13	174	0,20	60.336	2,95	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	128.470	7,31	3.934	7,76	5.491	3,56	3.925	4,60	141.820	6,93	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	68.300	3,89	90	0,18	147	0,10	112	0,13	68.649	3,35	Λ	Public administration & defence, compulsory social security
Εκπαίδευση	Μ	91.400	5,20	1.012	2,00	847	0,55	632	0,74	93.891	4,59	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	79.411	4,52	791	1,56	1.180	0,77	728	0,85	82.110	4,01	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	100.163	5,70	1.397	2,75	3.448	2,24	2.182	2,56	107.190	5,24	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.723	0,33	3.555	7,01	2.718	1,76	9.325	10,93	21.321	1,04	Ο	Private households with employed persons
Εξωδίκιοι οργανισμοί	Π	1.207	0,07	71	0,14	15	0,01	73	0,09	1.366	0,07	Π	Extra territorial organizations & bodies
Άγνωστο		12.564	0,72	336	0,66	722	0,47	609	0,71	14.231	0,70		Unknown
Σύνολο		**1.756.774**	**100,00**	**50.716**	**100,00**	**154.240**	**100,00**	**85.329**	**100,00**	**2.047.059**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 21,23%, στις μεταποιητικές το 17,93%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 22,62%, στις μεταποιητικές 18,07%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στα ξενοδοχεία και εστιατόρια 26,87% και στις κατασκευές 20%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 45,62% και στα ξενοδοχεία και εστιατόρια 16,40%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές βιομηχανίες 24,25% και στις κατασκευές 20,69% .

COMMENTS ON EMPLOYER'S ECONOMIC ACTIVITY

1) 21,23% of Insured Population is empoloyed in 'Wholesale & retail trade' and 17,93% in 'Manufacturing'
2) 22,62% of Insured Greeks is employed in 'Wholesale & retail trade', 18,07% in 'Manufacturing'
3) 26,87% of insured EU citizens is employed in 'Hotels & restaurants', 20% in 'Manufacturing'
4) 45,62% of insured Albanians is employed in 'Constructions'and 16,40% in ''Hotels & restaurants''
5) 24,25% of the insured foreigner workers is employed in 'Manufacturing' and 20,69% in 'Constructions'

ΔΕΚΕΜΒΡΙΟΣ 2007 / DECEMBER 2007

ΠΙΝΑΚΑΣ 0.6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 0.6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. (24) EU (24) Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	6.157	0,37	67	0,18	499	0,37	233	0,30	6.956	0,36	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.287	0,14	46	0,12	95	0,07	115	0,15	2.543	0,13	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.457	0,45	105	0,27	321	0,24	195	0,25	8.078	0,42	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	307.057	18,39	6.609	17,30	22.071	16,49	20.519	26,40	356.256	18,56	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	15.233	0,91	23	0,06	101	0,08	44	0,06	15.401	0,80	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	152.046	9,11	9.334	24,44	62.089	46,38	15.544	20,00	239.013	12,45	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	396.037	23,72	5.260	13,77	18.036	13,47	12.843	16,52	432.176	22,52	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	98.769	5,92	5.132	13,44	13.380	10,00	8.823	11,35	126.104	6,57	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	126.326	7,57	1.335	3,50	2.382	1,78	2.131	2,74	132.174	6,89	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	64.847	3,88	398	1,04	264	0,20	180	0,23	65.689	3,42	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	127.672	7,65	2.918	7,64	5.608	4,19	3.853	4,96	140.051	7,30	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	62.282	3,73	80	0,21	143	0,11	109	0,14	62.614	3,26	Λ	Public administration & defence, compulsory socpal security
Εκπαίδευση	Μ	112.055	6,71	1.429	3,74	988	0,74	795	1,02	115.267	6,01	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	81.492	4,88	765	2,00	1.274	0,95	746	0,96	84.277	4,39	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	90.703	5,43	1.188	3,11	3.312	2,47	2.022	2,60	97.225	5,07	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.490	0,33	3.147	8,24	2.583	1,93	8.903	11,45	20.123	1,05	Ο	Private households with employed persons
Ετερόδικοι οργανισμοί	Π	1.149	0,07	74	0,19	12	0,01	71	0,09	1.306	0,07	Π	Extra territorial organizations & bodies
Άγνωστο		12.419	0,74	283	0,74	706	0,53	606	0,78	14.014	0,73		Unknown
Σύνολο		**1.669.478**	**100,00**	**38.193**	**100,00**	**133.864**	**100,00**	**77.732**	**100,00**	**1.919.267**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός * The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ
1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 22,52%, στις μεταποιητικές το 18,56%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 23,72%, στις μεταποιητικές 18,39%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στις κατασκευές 24,44% και στις μεταποιητικές 17,30%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 46,38% και στις μεταποιητικές 16,49%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές βιομηχανίες 26,40% και στις κατασκευές 20% .

COMMENTS ON EMPLOYER'S ECONOMIC ACTIVITY
1) 22,52% of Insured Population is empoloyed in 'Wholesale & retail trade' and 18,56% in 'Manufacturing'
2) 23,72% of Insured Greeks is employed in 'Wholesale & retail trade', 18,39% in 'Manufacturing'
3) 24,44% of insured EU citizens is employed in 'Constructions', 17,30% in 'Manufacturing'
4) 46,38% of insured Albanians is employed in 'Constructions'and 16,49% in 'Manufacturing'
5) 26,40% of the insured foreigner workers is employed in 'Manufacturing' and 20% in 'Constructions'

ΙΟΥΝΙΟΣ 2008 / JUNE 2008

ΠΙΝΑΚΑΣ 0.6 : ΚΟΙΝΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΟΙΚΟΔΟΜΟΤΕΧΝΙΚΑ ΕΡΓΑ
ΚΑΤΑΝΟΜΗ ΑΣΦΑΛΙΣΜΕΝΩΝ* ΑΝΑ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΚΑΙ ΥΠΗΚΟΟΤΗΤΑ
TABLE 0.6 : ENTERPRISES & CONSTRUCTIONS
DISTRIBUTION OF INSURED POPULATION BY ECONOMIC ACTIVITY & NATIONALITY

Οικονομική δραστηριότητα		Ελλάδα Greece	%	Χώρες Ε.Ε. EU Countries	%	Αλβανία Albania	%	Υπόλοιποι Others	%	Σύνολο Total	%		Economic activity
Γεωργία, κτηνοτροφία, θήρα και σ.β.δ.	Α	5.523	0,31	88	0,17	681	0,42	246	0,27	6.538	0,31	Α	Agriculture, hunting and related service activities
Αλιεία	Β	2.344	0,13	53	0,10	90	0,06	104	0,11	2.591	0,12	Β	Fishing
Ορυχεία και Λατομεία	Γ	7.693	0,43	110	0,21	340	0,21	204	0,23	8.347	0,40	Γ	Mining & quarrying
Μεταποιητικές Βιομηχανίες	Δ	316.354	17,76	7.140	13,59	24.240	15,11	22.878	25,29	370.612	17,78	Δ	Manufacturing
Παροχή ηλ. ρεύματος φυσικού αερίου και νερού	Ε	16.264	0,91	36	0,07	127	0,08	62	0,07	16.489	0,79	Ε	Electricity, gas and water supply
Κατασκευές	ΣΤ	153.813	8,64	9.742	18,55	66.233	41,30	16.551	18,30	246.339	11,82	ΣΤ	Constructions
Χονδρικό και Λιανικό εμπόριο	Ζ	410.892	23,07	6.563	12,50	20.494	12,78	14.728	16,28	452.677	21,72	Ζ	Wholesale & retail trade
Ξενοδοχεία και Εστιατόρια	Η	167.010	9,38	14.912	28,39	28.803	17,96	14.918	16,49	225.643	10,83	Η	Hotels & restaurants
Μεταφορές, αποθήκευση και επικοινωνίες	Θ	130.995	7,36	2.317	4,41	2.721	1,70	2.686	2,97	138.719	6,66	Θ	Transport storrage & communications
Ενδιάμεσοι χρηματοπιστωτικοί οργανισμοί	Ι	65.825	3,70	419	0,80	278	0,17	199	0,22	66.721	3,20	Ι	Financial intermedation
Διαχείριση ακίνητης περιουσίας	Κ	135.273	7,60	4.253	8,10	6.461	4,03	4.406	4,87	150.393	7,22	Κ	Real estate, renting and business activities
Δημόσια διοίκηση και άμυνα	Λ	75.106	4,22	112	0,21	175	0,11	131	0,14	75.524	3,62	Λ	Public administration & defence, compulsory social security
Εκπαίδευση	Μ	95.306	5,35	1.091	2,08	960	0,60	662	0,73	98.019	4,70	Μ	Education
Υγεία και κοινωνική μέριμνα	Ν	84.616	4,75	833	1,59	1.461	0,91	807	0,89	87.717	4,21	Ν	Health & social work
Άλλες δραστηριότητες	Ξ	94.557	5,31	1.524	2,90	3.870	2,41	2.317	2,56	102.268	4,91	Ξ	Other community, social and personal service activities
Ιδιωτικά νοικοκυριά	Ο	5.342	0,30	2.912	5,54	2.616	1,63	8.795	9,72	19.665	0,94	Ο	Private households with employed persons
Εξωεδαφικοί οργανισμοί	Π	1.153	0,06	70	0,13	13	0,01	76	0,08	1.312	0,06	Π	Extra territorial organizations & bodies
Άγνωστο		12.856	0,72	349	0,66	824	0,51	694	0,77	14.723	0,71		Unknown
Σύνολο		**1.780.922**	**100,00**	**52.524**	**100,00**	**160.387**	**100,00**	**90.464**	**100,00**	**2.084.297**	**100,00**		**Total**

* Ο αριθμός των ασφαλισμένων είναι διακριτός

* The number of insured individuals is distinct

ΠΑΡΑΤΗΡΗΣΕΙΣ ΩΣ ΠΡΟΣ ΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΕΡΓΟΔΟΤΗ

1. Στο Σύνολο των ασφαλισμένων απασχολούνται στο χονδρ. & λιαν. εμπόριο 21,72%, στις μεταποιητικές το 17,78%
2. Οι Έλληνες απασχολούνται στο χονδρικό & λιανικό εμπόριο 23,07%, στις μεταποιητικές 17,76%
3. Οι Υπήκοοι των άλλων χωρών της Ε.Ε απασχολούνται στα ξενοδοχεία και εστιατόρια 28,39% στις κατασκευές 18,55%.
4. Οι Αλβανοί απασχολούνται στις κατασκευές 41,30% και στα ξενοδοχεία και εστιατόρια 17,96%.
5. Οι Υπόλοιποι αλλοδαποί εργαζόμενοι απασχολούνται στις μεταποιητικές βιομηχανίες 25,29% και στις κατασκευές 18,30% .

COMMENTS ON EMPLOYER'S ECONOMIC ACTIVITY

1) 21,72% of Insured Population is empoloyed in 'Wholesale & retail trade' and 17,78% in 'Manufacturing'
2) 23,07% of Insured Greeks is employed in 'Wholesale & retail trade', 17,76% in 'Manufacturing'
3) 28,39% of insured EU citizens is employed in 'Hotels & restaurants', 18,55% in 'Manufacturing'
4) 41,30% of insured Albanians is employed in 'Constructions'and 17,96% in 'Hotels & restaurants'
5) 25,29% of the insured foreigner workers is employed in 'Manufacturing' and 18,30% in 'Constructions'

Γ. Κώδικας Προγραμμάτων που υλοποιήθηκαν στο Matlab

Γ.1 Κώδικας του αρχείου change.m

Αυτό το m file μετατρέπει τους κωδικούς οικονομικής δραστηριότητας στις 17 βασικές κατηγορίες, όπως εμφανίζονται στα εξαμηνιαία στατιστικά δελτία του ΙΚΑ και στο ΣΤΑΚΟΔ 2003.

```
% programma poy metatrepei toys kodikous stakod-2003 opws 01,02,05,...
% sth megalyterh kathgoria kodikwn toy stakod-2003 opws A=1,B=2,...
(17 kathgories)
len=length(Poik); % Poik o arxikos pinakas
Pfinal=zeros(len,1);% Pfinal o telikos pinakas

for n=1:len,
    if Poik(n)<=4
        Pfinal(n)=1;  %  A
    elseif Poik(n)<=9
        Pfinal(n)=2;  %  B
    elseif Poik(n)<=14
        Pfinal(n)=3;  % Gamma
    elseif Poik(n)<=39
        Pfinal(n)=4;  %  Delta
    elseif Poik(n)<=44
        Pfinal(n)=5;  % Epsilon
    elseif Poik(n)<=49
        Pfinal(n)=6;  % ST
    elseif Poik(n)<=54
        Pfinal(n)=7; % Z
    elseif Poik(n)<=59
        Pfinal(n)=8; % H
    elseif Poik(n)<=64
        Pfinal(n)=9;  % Theta
    elseif Poik(n)<=69
        Pfinal(n)=10;  %  I
    elseif Poik(n)<=74
        Pfinal(n)=11;  %  K
    elseif Poik(n)<=79
        Pfinal(n)=12;  % Lambda
    elseif Poik(n)<=84
        Pfinal(n)=13;  %  M
    elseif Poik(n)<=89
        Pfinal(n)=14;  %  N
    elseif Poik(n)<=94
        Pfinal(n)=15;  %  ksi
    elseif Poik(n)<=98
        Pfinal(n)=16;  %  O
    else
        Pfinal(n)=17;  % Pi
    end;
end;
```

Γ.2 Κώδικας του αρχείου statist.m

Το πρόγραμμα αυτό υπολογίζει τις μεταβολές της οικονομικής δραστηριότητας στη διάρκεια του χρόνου για 9 εξάμηνα με δεδομένο το 2ο εξάμηνο του 2007 και τα δημοσιευμένα εξαμηνιαία στατιστικά δελτία του ΙΚΑ.

```
% PA= einai o pinakas me ta pososta ths arxikhs xronikhs periodou
% PB= einai o pinakas me ta pososta ths telikhs xronikhs periodou
% Pold= einai o pinakas me tis 65484 eggrafes tvn oikonomikwn
drasthriothtwn
% PE= einai o pinakas twn 3 ethnothtwn EU, AL, OT (65484 eggrafes).
Diavazetai mono to prwto gramma E, A, O
Pnew=Pold; % Pnew= einai o kainourios zhtoumenos pinakas
PD=PB-PA; len=length(Pnew);
SUMeu=[];SUMal=[];SUMot=[];% kathe pinakas SUM exei sthn prwth sthlh
mono tis oikonomikes drasthriothtes (1...17)
%pou afxanoun kai sth deuterh sthlh ta athroistika pososta ayths ths
afxhshs.

Seu=0;Sal=0;Sot=0; % athroistes twn thetikwn pososwn
Peu=[];Pal=[];Pot=[];% help matrices gia th deyterh sthlh twn SUMeu,
SUMal, SUMot
for k=1:17,
    if PD(k,1)>=0
        Seu=Seu+PD(k,1);
        SUMeu=[SUMeu;k]; % edw ftiaxnetai h prwth sthlh
        Peu=[Peu;Seu];% edw ftiaxnetai h deyterh sthlh me ta
athroistika pososta
    end;
    if PD(k,2)>=0
        Sal=Sal+PD(k,2);
        SUMal=[SUMal;k]; % edw ftiaxnetai h prwth sthlh
        Pal=[Pal;Sal];% edw ftiaxnetai h deyterh sthlh me ta
athroistika pososta
    end;
    if PD(k,3)>=0
        Sot=Sot+PD(k,3);
        SUMot=[SUMot;k]; % edw ftiaxnetai h prwth sthlh
        Pot=[Pot;Sot];% edw ftiaxnetai h deyterh sthlh me ta
athroistika pososta
    end;
end;

SUMeu=[SUMeu,Peu/Seu];SUMal=[SUMal,Pal/Sal];SUMot=[SUMot,Pot/Sot];

len1=length(SUMeu);len2=length(SUMal);len3=length(SUMot);


for n=1:len,
    if PE(n)=='E'
        if PD(Pold(n),1)<0
            x=rand(1); % Prwth klhrwsh gia to ean tha allaxei timh
            if x<=(-PD(Pold(n),1)/PA(Pold(n),1))
                y=rand(1); % Deyterh klhrwsh gia to pws tha allaxei
                for k=1:len1,
                    if (y-SUMeu(k,2))<=0
                        Pnew(n)=SUMeu(k,1);
                        break;
```

```
                    end;
                end; % for loop k
            end;
        end;

    elseif PE(n)=='A'
        if PD(Pold(n),2)<0
            x=rand(1); % Prwth klhrwsh gia to ean tha allaxei timh
            if x<=(-PD(Pold(n),2)/PA(Pold(n),2))
                y=rand(1); % Deyterh klhrwsh gia to pws tha allaxei
                for k=1:len2,
                    if (y-SUMal(k,2))<=0
                        Pnew(n)=SUMal(k,1);
                        break;
                    end;
                end; % for loop k
            end;
        end;

    else
        if PD(Pold(n),3)<0
            x=rand(1); % Prwth klhrwsh gia to ean tha allaxei timh
            if x<=(-PD(Pold(n),3)/PA(Pold(n),3))
                y=rand(1); % Deyterh klhrwsh gia to pws tha allaxei
                for k=1:len3,
                    if (y-SUMot(k,2))<=0
                        Pnew(n)=SUMot(k,1);
                        break;
                    end;
                end; % for loop k
            end;
        end;


    end;

end;
```

Γ.3 Κώδικας του αρχείου apografh.m

Το πρόγραμμα αυτό παράγει το έτος απογραφής με βάση τις παραδοχές που ορίστηκαν στο κείμενο.

```
len=length(Pbirth);
Papogr=Pbirth;

for n=1:len,
    if Pbirth(n)==1987
        Papogr(n)=2004;
    elseif Pbirth(n)>=1975
        Papogr(n)=round(Pbirth(n)+16.5+rand(1)*(2004.5-16.5-
Pbirth(n)));
    else  % for year<=1974
        if PE(n)=='E'    % EU
            x=rand(1);
            if x<=.45
                Papogr(n)=round(1990.5+rand(1)*7);
            else
                Papogr(n)=round(1997.5+rand(1)*7);
            end;
        elseif PE(n)=='A'  % AL
            x=rand(1);
            if x<=.7
                Papogr(n)=round(1990.5+rand(1)*6);
            elseif x<=.9
                Papogr(n)=round(1996.5+rand(1)*5);
            else
                Papogr(n)=round(2001.5+rand(1)*3);
            end;
        else             % OT
            x=rand(1);
            if x<=.4
                Papogr(n)=round(1990.5+rand(1)*7);
            else
                Papogr(n)=round(1997.5+rand(1)*7);
            end;
        end;
    end; % IF
end; % FOR
```